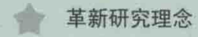
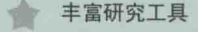


格致方法 · 定量研究系列　吴晓刚　主编

空间回归模型

[美] 迈克尔 · D.沃德（Michael D. Ward）
克里斯蒂安 · 格里蒂奇（Kristian Skrede Gleditsch）著
宋曦 译


革新研究理念


丰富研究工具

最权威、最前沿的定量研究方法指南

11

格致出版社　上海人民出版社

格致方法·定量研究系列　吴晓刚　主编

空间回归模型

[美] 迈克尔·D.沃德(Michael D.Ward)
克里斯蒂安·格里蒂奇(Kristian Skrede Gleditsch) 著
宋　曦 译

SAGE Publications ,Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

空间回归模型/(美)沃德(Ward, M. D.),(美)格里蒂奇(Gleditsch, K. S.)著;宋曦译. —上海:格致出版社:上海人民出版社,2016

(格致方法·定量研究系列)

ISBN 978-7-5432-2615-9

Ⅰ.①空… Ⅱ.①沃… ②格… ③宋… Ⅲ.①回归分析-研究 Ⅳ.①0212.1

中国版本图书馆CIP数据核字(2016)第062894号

责任编辑 王亚丽

格致方法·定量研究系列

空间回归模型

[美]迈克尔·D.沃德
克里斯蒂安·格里蒂奇 著

宋曦 译

出 版	世纪出版股份有限公司 格致出版社 世纪出版集团 上海人民出版社 (200001 上海福建中路193号 www.ewen.co)	印 刷	浙江临安曙光印务有限公司
	编辑部热线 021-63914988 市场部热线 021-63914081 www.hibooks.cn	开 本	920×1168 1/32
		印 张	5
		字 数	96,000
		版 次	2016年4月第1版
发 行	上海世纪出版股份有限公司发行中心	印 次	2016年4月第1次印刷



ISBN 978-7-5432-2615-9/C·145 定价:25.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的 SAGE 出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于 2011 年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于 2012 年以单行本的形式再次出版发行，共 37 本。我们衷心感谢广大读者的支持和建议。

随着与 SAGE 出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

2003 年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”（Statistics for Social Science）为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课（事实上，大部分硕士生在修完第一门课后都会继续选修第二门课）。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复，但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂，与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在"格致方法 · 定量研究系列"丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

社会科学定量研究的很多方法都是基于对个体行为的分析。我们通常将个体行为作为结果变量,然后将其表示成由一组解释变量构成的方程,最常见的就是回归方程的形式。社会科学理论主要就是描述这种结果变量和解释变量之间是如何关联的。当然,这种分析不仅停留在个体或者微观层面。有时,研究者也会分析集体层面的数据,比如街道、社区、公司、城市、县、州以及国家。但是分析的逻辑仍然相同。我们试图构造因变量和某些自变量(诸如个人、街坊、社区、公司、城市、县、州和国家)之间的关联,不管这种关联是否为因果关系。当我们这样做的时候,其实已经暗示了观测值的地理或者空间位置并不起作用。尽管研究者也会经常用虚拟变量将处于同一个地区的观测值合成一组,但这样做并没有考虑除了空间以外的其他相似性。比如,数据分析者常用虚拟变量来划分个体是来自美国南部还是其他地区。这样做是为了控制一些特有的文化特征,而不是出于对回归中空间依赖关系的考虑。《空间回归模型》一书将不遗余力地解决线性回归分析中空间依赖关系的相关问题。

确切地说，我们将结果变量 y_i 对解释变量的向量 x_i 做 OLS 回归：

$$y_i = x_i\beta + \varepsilon_i$$

这里 β 包括被估参数的一组向量，ε_i 是服从独立同分布假设的随机误差。在经典线性回归中，假设分布为正态分布。当存在空间（或其他）依赖关系时，ε_i 将不再相互独立，并且导致 β 被低估，从而影响假设检验的正确性。

尽管对空间问题的考虑可以追溯到早期在地图制作和调查方面的尝试，但是现代空间回归是直到最近几十年随着统计知识和计算能力的提高才出现的。本书作者向读者介绍了两种应用最广泛的空间回归模型：空间性定距因变量和空间性误差模型，此外还补充了空间分析中的疑难问题。尽管读者的分析单位可能与书中作者的分析单位并不相同，然而书中大量直观的例子仍然能够为读者提供有益的思考。

廖福挺

前 言

空间观念能够为社会科学研究作出巨大贡献。本书试图填补这一领域的空缺，为社会科学家完备地介绍如何将空间依赖性的分析纳入回归框架，让更多对社会科学中空间应用问题感兴趣的读者能够读懂此书。尽管当前已经有非常详尽的空间统计学调查，但是它们中的大部分都过于深奥，并且都假设读者已经具有丰富的高级统计和概率论知识(Banerjee、Carlin & Gelfand, 2004; Cressie, 1993; Getis & Boots, 1978; Haining, 2003; Ripley, 1981、1988; Schabenberger & Gotway, 2005)。此外，这些调查中的大部分都是自然科学的主题或者该方面的应用，而并不为社会科学研究者所熟知。我们假定此书读者仅仅了解社会科学研究中广泛应用的经典回归模型，同时他们对数据中可能存在的空间依赖性问题感兴趣。在某些部分，虽然我们会用到矩阵的表示形式，但同时也会用非数学语言对其进行详细解释。我们会用到免费且应用广泛的 R 计算平台(R Development Core Team, 2004)来演示如何使用这些方法，同时还提供了一段编码，对例子进行解释。如果读者对 R 的了解达到了达尔加

德一书(Dalgaard，2002)所介绍的程度，将非常有助于理解本书的内容。尽管其他的程序和方法也可用于分析空间数据，但是本书中将不予以介绍，不过我们会以附录的形式提供一些简要的细节内容。

本书的面世离不开外界的帮助。首先，我们要感谢家庭成员对于我们紧张工作的体谅。同时我们要感谢西班牙巴塞罗那经济分析研究所主任琼·埃斯特班(Joan Esteban)，她给予我们热情慷慨的欢迎和巨大的支持，帮助我们完成了拖延很久的第一稿。迈克尔·沃德还要感谢统计与社会科学中心主任埃德里安·拉夫特里在他访问期间提供的帮助。同时，沃德还要感谢华盛顿大学艺术与科学学院前院长和现任迈阿密大学校长大卫·霍奇以及华盛顿大学政治科学系主任史蒂夫·马耶斯基的帮助。克里斯蒂安·格里蒂奇得到了来自国家科学基金会的赞助以及来自加泰罗尼亚政府和加州大学的加斯帕·波多拉的旅行资助。

华盛顿大学、加州大学圣地亚哥分校、埃塞克斯大学等地的前任或现任同仁们引导并激发了我们在依赖性数据上的兴趣。我们同样要感谢约翰·阿尔齐斯特、克里斯·巴基、凯尔·比尔兹利、纳撒尼尔·贝克、罗杰·比万德、曹汛、肖娜·费希尔、詹姆士·勒萨热、林泽民、迈克尔·曼格尔、阿西姆·普拉卡什、安德里亚·鲁杰里、伊德里·萨里希安、迈克尔·锡恩、克里斯多夫·沃德、丹尼尔·沃德、安东·韦斯特维尔德三世和艾里克·韦伯尔斯给予的深刻见解和有益的讨论。迈克尔·D. 沃德要感谢他以前在科罗拉多大学行为科学研究所的两位邻座安德鲁·科比和约翰·奥洛林对他在地理方面兴趣的影响。克里斯蒂安·格里蒂奇感谢

基德龙的《国家的世界地图集》(*State of the World Atlas*)一书对他在地理和社会科学方面兴趣的影响。

在20世纪，迈克尔·刘易斯—贝克鼓励我们从事这个项目。他是一个有耐心的人。

迈克尔·D.沃德、克里斯蒂安·格里蒂奇

目 录

第1章

导　论

第1节 | 交互作用与社会科学

社会科学研究者总是对各种情况下不同行动的中间人(例如:个人、政治团体、群体和国家)之间的交往感兴趣。在很多情况下,个体行动者的行动动力和结果并不完全取决于个人特征,而是取决于社会结构及个人的社会位置,以及个体与个体之间的互动。即便是像流感这么平常的一件事情,也有社会因素的作用在里面,因为流感的传播也需要通过社会交往。比如想要预测一个人是否有可能染上鼻病毒,我们就需要考察最近周围是否有异常发生,同时这个人和感染鼻病毒的人是否接触过。有一些疾病是通过接触传染的,也就是感染者在和其他人的交往当中传播疾病。显然,不同类型的交往会导致不同的疾病。在20世纪70年代末期,艾滋病在美国的传播方式被指认为来自加拿大航空公司的单身服务员(Watt, 2003),但这已经被确认为误传。

奇怪的是,交往的作用以及交往结构在社会科学研究中却几乎完全被忽略。比如说,在投票数这件事上,投票率差异以前都是通过个体特征,诸如教育高低或者对政治行为的重视程度来解释的。然而,社会交往以及个体之间的相互联系是与个人特征同样重要的因素。例如,动员投票的电话平均会使投票率变动6个百分点(±3%)(Imai, 2005)。类似

的，个人和教堂、工会等组织有联系，也会增加投票率。贝别克和哈克菲尔德(Baybeck & Huckfeldt, 2002)就发现：即便是在分散化的网络当中，间隔较远的个人之间也更少有频繁的交往。这种研究通常是一些例外，而不是惯例。大部分投票率研究仍然假定所有投票者的决策是相互独立的。

显然，在流感的例子中，将个体之间看做毫无关联，是一种明显不合逻辑的做法。有一些人可能免疫力更差，所以更可能在流行病发生时得病。然而，我们不可能在不知道其他人是否染病的情况下，仅仅通过个人特征就预测出个人患流感的几率。又比如，父母通常在收入、睡眠时间、吸烟史上都与子女不相同。然而，一方有某种习惯，另外一方通常也会受到影响。社会关系模型的发展来源于心理学家对分离群体和个体间独立效应和交互效应的兴趣，同时他们试图将这种依赖关系用模型表示出来(参见 Kenny, 1981; Malloy & Kenny, 1986)。

在本书中，我们考察了空间分析的视角如何帮助研究者处理观测值之间的依赖关系及处理空间聚类现象。我们尤其关注两类含有空间因变量的回归模型。第一类是关注含有空间滞后因变量(spatially lagged dependent variable)的情形。第二类是关注空间性误差(spatial error)。我们也意识到：空间性模型其实存在很多有趣的视角，但本书除了介绍空间滞后因变量和空间性误差，将不考察其他问题。尽管这些视角对于社会科学的经验研究大有裨益，然而到目前为止，很多文献仍然没有对此引起重视。这些模型使得我们可以考察一个观测值对其他相近观测值的影响。当然，我们相信这种重要性不仅体现在基本原理方面，最简单的道理其实

是因为很多社会现象都是空间性"聚集"的。这些空间上排列的数据,既包括地理上观测到的个体位置,也包括在某个地理区域里的地区性数据。后一种数据类型称之为地区(area)或晶格(lattice)数据,而前者称之为点状(point)数据。在本书中,我们将重点放在地区性数据上,这种数据通常用于处理县、州、省、国家等个体单位。[1]

社会科学中空间聚类现象非常普遍。投票的地区聚类问题被认为在美国人的政治行为上起到了重要作用。政治分化与经济与种族上的分化是联系在一起的。正因为如此,投票率模型才不得不将各种分化的空间聚类效应考虑在内(West, 2005)。相似的例子在比较政治学、社会学和经济学研究当中也可以找到。比如,在有关中央银行不同政策的影响的研究中,就有人曾考察过这些政策选择和中央银行以及银行家们的偏好之间的独立性。一种广泛的看法认为:不管中央银行如何不受国家主管机构的影响,它仍然受到各种各样地方情况的约束。因此,即便中央银行和国家主管机构之间互不干涉,中央银行的政策相互之间也是独立的吗(Adolph, 2004; Franzese, 1999)? 默多克、桑德勒和萨金特(Murdoch、Sandler & Sargent, 1997)研究过20世纪80年代欧洲排放硫化物和一氧化二氮的行为,其自愿和非自愿决定之间的相互影响关系。由于污染者们在空间上是分散的,并不受国界的限制,空间分析技术将有助于强调污染的外溢效应以及约束履行上的相互影响问题。在跨国研究中,不平等和贫困问题被认为是交织在一起的。在越贫穷的国家,财富和收入分配越不均衡。当前研究也发现,贪污通常是贫困产生的结果,同时也可能成为贫困产生的原因。然而,研究也

发现:收入不平等可能会增加贪污的程度,甚至大于对贫困的影响。一种更复杂的可能的情况是,财富和贪污在空间上具有聚类效应。空间分析就可以帮助我们解决这个难题。尤和卡格拉姆(You & Khargram, 2005)的当前研究就是按照这种思路进行的。最后一点,组织形式的扩展可能也是遵循相同的方式——比如政策效仿。霍姆斯(Holmes, 2006)就利用空间模型探讨过工会组织的蔓延问题。

简言之,社会科学中有无数研究数据都是按照一定的空间形式组织起来的,不管它们的分析单位是县、市、州、国家还是公司。通常,这些分析单位的特征都是高度聚类的,尤其是空间层面的地区聚类。在很多应用中,合乎逻辑的假设都应该包括观测值之间的相互关联性。在实际操作方面,这些聚类通常都被忽略不计或者被当做一种干扰。忽略这种关联性将会极大地影响我们在研究中建立有意义的推论。空间分析不仅为减少这种代价提供了一种方法,同时空间信息将有助于揭示社会过程之间是如何联系起来的。下面我们就将介绍在社会科学的一个重要分支中应用这种分析的简单例子——关于民主制度扩散的研究。

第2节 | 世界各国的民主

在引入空间讨论之前，我们先举一个数据中各观测值之间相互不独立的例子。社会科学家们很早就乐于解释为什么有的国家采用民主政治而有的却不是如此。早期由李普赛特提出的一项颇具影响力的观点认为：民主制度是具有社会必要条件的。其中之一就是较高的平均收入，李普赛特注意到，“在更加民主的国家……平均财富也更高”(Lipset, 1959:75)。在过去的40多年里，这种观点成为比较分析领域的奠基石，它表明平均收入更高的国家更可能建立民主体制。表1.1提供了一个数据简表，其中列出了2002年世界70多个国家的人均国内生产总值(GDP)和民主水平。我们对民主的测量来自POLITY指数，它将国家按照一系列制度标准划分成不同的类别。在这一指数中，−10代表最不民主的社会，10代表最民主的社会。格里蒂奇和沃德(Gleditsch & Ward, 1997)提供了建构这种指数的更详细信息。我们在表1.1中将各国按照人均国内生产值(GDP)和民主程度高低排序，以便于找出变量之间的简单关系。正如我们所见，一些富裕国家，比如丹麦，的确是民主国家；同时低收入国家，比如塞拉利昂和朝鲜，就是专制国家。有趣的是，李普赛特曾提出，1959年，澳大利亚、比利时、加拿大、丹麦、爱尔

表 1.1 2002 年 GDP 数值

国家	民主	GDP	国家	民主	GDP
几内亚	−1	51	伊朗	3	1776
埃塞俄比亚	1	114	马其顿	6	1801
布隆迪	0	120	纳米比亚	6	1870
扎伊尔	0	135	罗马尼亚	8	1941
塞拉利昂	−10	172	阿尔及利亚	−3	2036
厄立特里亚	−7	175	波斯尼亚和黑塞哥维亚	0	2108
马拉维	5	178	泰国	9	2215
伊拉克	−9	181	苏里南	9	2224
几内亚比绍	5	187	危地马拉	8	2257
利比亚	0	194	俄罗斯	7	2279
卢旺达	−4	216	厄瓜多尔	6	2305
莫桑比克	6	217	秘鲁	9	2306
塔吉克斯坦	−1	221	哥伦比亚	7	2342
尼日尔	4	247	约旦	−2	2375
尼泊尔	6	276	斐济	5	2397
布基纳法索	0	315	突尼斯	−4	2436
乍得	−2	317	萨尔瓦多	7	2486
乌干达	−4	320	南非	9	2607
坦桑尼亚	2	330	多米尼加共和国	8	2745
中非	5	333			
…	…	…	…	…	…
土库曼斯坦	−9	1241	加拿大	10	25139
摩洛哥	−6	1300	芬兰	10	26235
刚果	−5	1303	澳大利亚	10	26304
吉布提	2	1313	荷兰	10	27059
白俄罗斯	−7	1359	瑞典	10	27497
			英国	10	27650
斯威士兰	−9	1412	日本	10	31731
阿尔巴尼亚	5	1416	阿联酋	−8	34436
叙利亚	−7	1417	卡塔尔	−10	36611
哈萨克斯坦	−6	1437	丹麦	10	37063
塞尔维亚	7	1573	瑞士	10	39769
埃及	−6	1602	美国	10	40180
缅甸	−7	1729	挪威	10	43895
保加利亚	9	1744	卢森堡	10	54255

注:GDP 值代表人均国内生产总值。本表有删节,全部数据可以在本册书网站上获取。

兰、卢森堡、荷兰、新西兰、挪威、瑞典、瑞士、英国和美国构成了欧洲、北美和南美各洲一系列的“稳定民主体制”。而当时不稳定的民主体制和独裁体制包括奥地利、芬兰、法国、前联邦德国、意大利和西班牙。如今这些国家也变成了民主体制，并且基本上稳定了。尽管这些个案同李普赛特的论调是一致的，但是，财富和民主之间是否存在更普遍的强烈相关关系？尽管印度的平均国民收入很低，但它却实行了民主体制；此外尽管印度近年来经历了高速的经济增长，但它仍然远远低于经济合作与发展组织（Organization for Economic Cooperation and Development，OECD）成员国的水平。同时，中东地区很多专制国家却拥有很高的收入水平，这也是与李普赛特的观点相违背的。为了更一般化地估计这种关系，我们需要一种更加系统性的比较分析。

自李普赛特（Lipset，1959）和以往其他学者的研究之后，有关民主的实证比较研究中通常都将民主作为人均 GDP 自然对数的线性函数。我们用 POLITY 分数估计一个国家的民主水平，在纳入人均 GDP 之后，将其表示为普通最小二乘法（OLS）回归：

$$\text{POLITYscore} = \beta_0 + \beta_1 \ln \text{GDPpercapita} + \varepsilon$$

该民主水平对于人均 GDP 的线性回归模型的估计结果见表 1.2。人均 GDP 自然对数的正向回归系数表明民主和收入之间是正向关系，但是如果考虑到变量的测量单位，估计得到的影响效果其实是较小的。

确切地说，此回归模型预测如果一个国家具有乌兹别克斯坦的人均 GDP（2002 年为 464 美元），那么这个国家的民

表 1.2 将民主视做人均 GDP 对数的线性方程的 OLS 估计

	$\hat{\beta}$	$SE(\hat{\beta})$	t Value
截距	−9.69	2.43	−3.99
人均 GDP 对数	1.69	0.31	5.36
$N = 158$			
Log likelihood ($df = 3$) $= -513.62$			
$F = 28.77$ ($df_1 = 1$, $df_2 = 156$)			

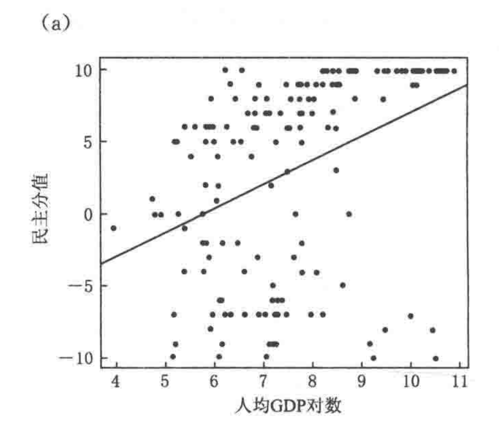
注:估计值来自于 POLITY 项目和世界银行 2002 年数据。

主得分将接近于 1。反过来,如果一个国家的人均 GDP 收入接近乌兹别克斯坦的两倍(1020 美元),该模型就可以预测相对应的民主得分约为 2。对于大多数分析者来说,在 POLITY 民主得分指数中,得分 1 和 2 是非常相似的。因此,即便收入水平上存在相对较大的差异,对民主水平的预测结果也不会相差太大,尽管人均 GDP 对数的估计系数在统计上是显著的。

图 1.1 表明 OLS 方程对贫穷国家民主水平的预测远远高于它们的实际水平。在贫穷国家中(比如乌兹别克斯坦),财富对民主潜在影响的估计效应不仅很小(即使人均 GDP 翻倍,对民主的影响作用也是很小的),而且这种效应还可能是被高估了。对这些残差的任何标准分析几乎都印证了该图给人的第一印象:这些残差看上去并没有“很好地分布”,因为在最高值和最低值附近存在观测值的两个峰值,这表明模型低估或者高估了实际的民主水平。

图 1.1 同时也表明围绕估计回归直线,或者说总体趋势上存在大量和呈规律的变化。但是这些残差的排列方式是由于观测值之间的相互关联造成的吗? 图 1.1(b)有力地表明残差并不呈正态分布,并且也不是一种单一模式;而是在

(a)



(b)

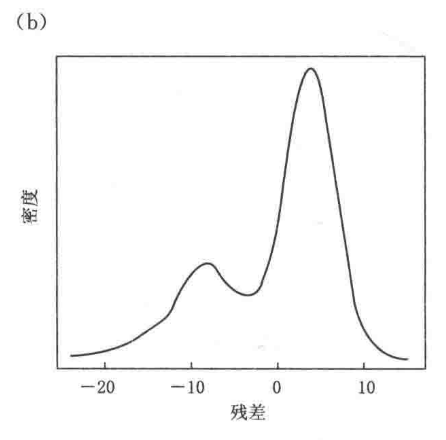


图 1.1 (a) 民主和人均 GDP 自然对数的散点图以及回归直线；(b) OLS 残差密度图。它们表明分布较低端存在独裁体制的“聚类”效应和分布较高端存在民主体制的“聚类”效应。

值−10 附近有一组负值，同时在 5 附近聚集着一组正值。因此，这个例子明确表明表 1.2 报告的 OLS 回归残差是有问题的，进一步回归得到的估计系数是否可信，也成为疑问。这些残差表明：该模型没有很好地把握民主和经济产出之间的关系，部分原因可能是因为数据之间的关联，即相似值之间的聚集作用。这可能是因为国家之间的相互影响所致。

第3节 | 空间依赖关系介绍

对于以上结果,一种可能的解释是:除了各个国家自身的特性以外,一个国家是否希望建立民主制度和周围国家是否已经建立民主制度是相互关联的。在冷战期间,苏联的干预就使得社会主义制度在很多东欧国家推行。此外,很多拉美国家发生的民主转型似乎也受到了其他国家社会进程的影响(参见 Gleditsch & Ward, 2007)。如果将表 1.1 的数据按照字母顺序排列,我们将很难判断是否存在一些相似体制国家所构成的区域,不同于我们从人均 GDP 中得到的预期类型。即便是按照国家的重要特征排列以后再进行比较,我们也需要经过仔细分析才能识别各种不同类型。

在很多情况下,检验可能存在的空间(以及类似空间的)聚集效应都是非常重要的,它将有助于我们发现是否因为表面的无关联而忽略了内在的社会互动关系。潜在且未观测到的聚集效应,可能会影响我们对于已有模型的理解,从而影响我们真正发现背后的实际过程。在讨论空间相关性之前,我们会先解释为什么这样做很重要。

即便分析者们只是想要比较均值和建立经典统计检验,例如均值差检验,如果数据中存在空间相关关系,那么这种做法也存在问题。假设有一个针对变量 y 的单样本的 t 检

验，如下：

$$t = \frac{1/n\sum_{i=1}^{n} y_i}{\sigma/\sqrt{n}}$$

如果邻近的观测值之间在时间上或者空间上相关（一阶序列相关），那么对于正向序列相关的值，它们的实际标准误将会偏大（对于负向值将会偏小）。研究者们对于不同时间点上观测值之间的序列相关问题一直比较谨慎，但是他们常常忽略的是，即便是在同一时间点上，不同观测值之间也可能出现同样的序列相关问题。通过方差的非调整估计得到的 t 值将比真实值更大。这将增加第一类错误（Type I）发生的可能性，即便是在空间自相关作用很小、观测值很多的情况下，也不例外。

简言之，由于观测值之间的空间序列相关（或其他原因），通过经典检验接受的假设结果将是有偏的，即便在检验结果不真实的情况下也是如此。假设数据在空间上是关联的，例如这种关联与观测值之间的距离成反比，ρ 代表一阶序列相关的空间相关系数。这种相关测量了相邻值之间在一些测量属性上的相似程度。这种相关导致均值的真实标准误将近似于：

$$\sigma_{\bar{y}} \approx \sqrt{\frac{1+\rho}{1-\rho}}\,\frac{\sigma}{\sqrt{n}}$$

一种简单的理解空间相关的影响的方法，是假设一个变量 y 有 n 个观测值：y_1，y_2，…，y_{n-1}，y_n。在很多情况下，我们都认为这些观测值之间相互独立，并且服从同一分布，一般是具有未知均值 μ 和方差 σ^2 的正态分布。对于 μ 的一般

估计值为：

$$\bar{y} = \sum_{i=1}^{n} y_i / n$$

由于假定观测值来自于正态分布，那么统计推断将基于 y 和 σ。95%的置信区间为 $\bar{y} \pm 1.96\sigma/\sqrt{n}$。如果 y_i 之间存在空间相关，也就是观测值 y_i 和 y_j 空间上隔得越近相似性越大，那么如同克里斯(Cressie, 1993:14)指出的一样，对于取值为正的 ρ，其协方差将为：

$$\mathrm{cov}(y_i, y_j) = \sigma^2 \times \rho^{|i-j|}$$

其方差为：

$$\mathrm{var}(\bar{y}) = n^{-2} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{cov}(y_i, y_j) \right\}$$

展开即为：

$$\mathrm{var}(\bar{y}) = \left\{ \frac{\sigma^2}{n} \right\} \left[1 + 2\left\{ \frac{\rho}{1-\rho} \right\} \left\{ 1 - \frac{1}{n} \right\} - 2\left\{ \frac{\rho}{1-\rho} \right\}^2 \frac{1-\rho^{n-1}}{n} \right]$$

其中的因子为：

$$\left[1 + 2\left\{ \frac{\rho}{1-\rho} \right\} \left\{ 1 - \frac{1}{n} \right\} - 2\left\{ \frac{\rho}{1-\rho} \right\}^2 \frac{1-\rho^{n-1}}{n} \right]$$

实质上是根据空间相关程度对观测值的数量打了折扣，并且这种作用不会随着样本量增大而消失。如果 $n = 10$、$\rho = 0.26$(按照克里斯的例子)，那么这种折扣效应约为 40%：10 个空间相关的观测值的精度相当于 6 个独立观测值。换句话说，这也表明当观测值之间存在空间正向相关的时候，忽略这种相关所得到的置信区间将远远窄于真实的情况。一般来说，忽视空间依赖性，将可能导致对于数据真实方差

的低估。因此，对于一个包含 158 个 GDP 观测值的样本，在正态分布假设下，其 95%的置信区间为$(1.96\times\sigma)/\sqrt{n}$，但是如果存在 0.65 的空间相关——上面例子中 GDP 的真实$\hat{\rho}$值——那么正确的置信区间将会接近于 4.22 而不是 1.96，几乎是原来的两倍。在民主发展程度的例子中，$\hat{\rho}$ 为0.47，那么 95%的置信区间为$(3.26\times\sigma)/\sqrt{n}$，几乎增大了 70%[2]。

如果空间相关具有不同的形式，那么就需要进行不同的具体调整，但是总的原则是如果空间相关为正，那么样本均值的精确性将降低。这通常将导致我们拒绝一个实际为真的零假设。此外，如果数据在空间上具有依赖关系（或者相互关联），那么基于独立同分布（iid）的假设进行的统计检验将变得很不可靠。斯卡本伯格和果特威（Schabenberger & Gotway，2005）详细叙述了在不同样本量和不同程度的自相关情况下最小二乘估计的过度变化。对于$\rho>0$，这种过度变化随着n的增大而增大，比如当$\rho=0.9$、样本量接近 50 的时候，这种过度变化将接近于 14.0。这里最重要的一点是，空间相关的数据将使基于 iid 假设的统计检验出现严重的问题，这也使得研究者因为标准检验低估了数据的变化而拒绝零假设。

第 4 节 | 将地图作为可视化数据

人们很擅长发现模式,即便是在没有模式存在的情况下,也是如此。通常这也就是统计的作用。然而,在具有探索性和启发性的模型中,尽可能多地了解数据是非常有用的。包含丰富信息的密集表格是传递大量信息的重要方式,但是这种做法比较慢。图形展示作为一种辅助方法,将有利于从视觉上很快发现模式的存在。然而,重要的是,这种图形技术应该用于提供研究现象的可能解释上。当前的研究已经阐述了仔细展示证据和定量材料的重要性,并且提供了黄金准则(Cleveland, 1993; Tufte, 1990、1992、1997; Wainer, 2004)。一个指导性的原则是展示方法应当与已有的解释之间存在密切关系。

一项有关伦敦 19 世纪中叶霍乱传播的经典研究就提供了一个基于地理解释的范例。该研究最早由约翰·斯诺(John Snow)提出,后被蒂夫特(Tufte, 1997)推广,最近被约翰逊(Johnson, 2006)进一步完善。斯诺证明:1854 年夏,伦敦霍乱爆发的原因在于家庭办公一族(以及其他人)的饮水来自于布罗德大街的水泵,而这些水受到来自于霍乱受难者墓场的污染。因此,靠近布罗德大街的水井就成为感染霍乱的潜在危险因素,同时这一研究也对否定霍乱的空气传播论

起到了重要作用。斯诺的伦敦地图也成为利用空间相关展现因果联系的重要例子。图 1.2 提供了家庭办公地区的经典地图,从中可以看出,霍乱所导致的死亡大多聚集在布罗德大街的水泵附近。



图 1.2 约翰·斯诺关于 1854 年夏伦敦家庭办公(Soho)区的霍乱死亡地图

在地图上标注阴影也是展现包含地理因素的发生过程的一种重要方式。我们的例子表明:邻近的国家之间会存在相互反馈,这将影响它们的政治体制和经济财富。图 1.3 的世界地图将 158 个国家在 2002 年的民主发展程度用阴影标示出来。这个地图告诉我们民主体制在相邻国家之间有聚集作用,同时在世界不同区域内的专制国家也是相互聚集

的。在该地图中阴影越深，表示民主发展程度越高。

只有那些具有最高民主得分的国家才被标记为黑色，例如法国被标记为第二深的阴影，这是因为它的民主得分为 9，这也反映了它的总统和国民议会之间相对独立。图 1.3 表明，民主制度和专制制度都存在很强的地理聚集效应。总的来说，大多数民主国家都位于西欧和美洲或者澳洲和大洋洲沿岸，而很多专制国家都位于非洲、中东和亚洲。

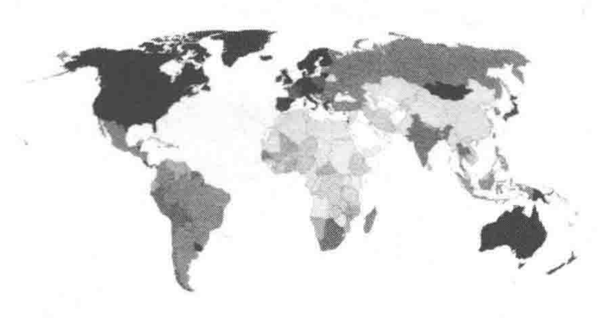


图 1.3 高民主水平国家用深色阴影表示，它们在世界范围内具有聚集趋势。类似的，更专制的国家在地理上也具有聚集趋势。民主水平用 POLITY IV 指标测量。

当然也存在例外，比如白俄罗斯，尽管周围大都是民主邻国，但它却依然保持着专制体制。相比之下，印度是一个民主国家，尽管周围邻国大多是非民主体制。

在拉丁美洲，尽管国家之间在人均 GDP 上存在巨大差异，但大部分国家在 2002 年都是民主国家。相对而言，尽管中东国家的人均 GDP 水平普遍比世界平均水平高，但其中大部分都是专制体制。事实上，地图上的这些特征告诉我们：民主和人均 GDP 都是在空间上聚集的。在很多情况下，可视化和地图化都有利于揭示数据结构，而这种结构在表格形式的数据中是不容易发现的。

图 1.4 展示了 2002 年人均 GDP(取对数)的聚集情况。富裕的国家用更深的阴影标出,而更贫困的国家用更浅的阴影。图 1.4 也反映了很强的聚集效应。北美和西欧都是富裕国家的聚集地区,而非洲则表明通常贫困国家的邻国也处于贫困之中。当然也存在例外,比如日本和澳大利亚总体来说都比它们的邻国要富裕很多。

将数据地图化将极大地帮助我们解释空间数据,但是好的展示也应该包括经验或者理论的解释。

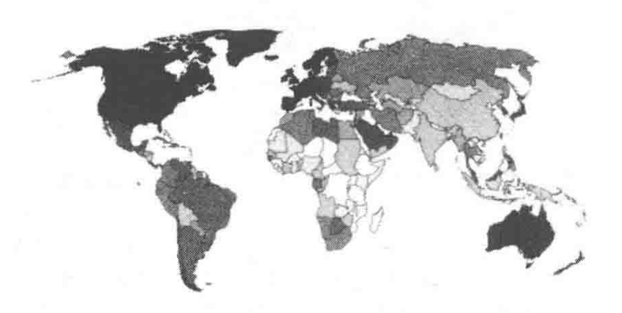


图 1.4 用更深的灰色表示的具有更高人均 GDP 水平的国家

第5节 | 空间依赖性和相关性测量

可惜，正如研究者可能忽略数据矩阵中某种模式一样，他们也可能在没有结构存在的情况下找出了某种结构。正因为如此，我们有必要寻找更正式的方式，来衡量观测值是否存在空间聚集，或者相互之间是否通过某种联系关联在一起。我们将在下一节转向正式的探索方法。

然而，解释这些关联需要我们已知哪些观测值有可能相互联系。对于包括 n 个元素的一组数据集，每个观测值 i 可能会和剩下的 $(n-i)$ 个可能的元素相关，但是，在实际中，我们通常假设某些相关关系或者联系将比其他的更重要。我们感兴趣的元素之间的网络或者结构通常必须在分析其他元素之间的关联关系之前就被明确指定。这里我们用到的技术通常起始于关联观测值之间关系的一幅图或者一个列表 L。基于种种原因，符合实际的做法是利用矩阵来表示观测值之间的关联性。比如，我们定义一个二元矩阵 C 来指定个体观测值之间的联系。如果两个观测值 i 和 j 被认为相互关联，那么输入值 $C_{ij}=1$，反之则 $C_{ij}=0$。

测量空间关联和相关的基本思路类似于向量内积，根据休伯特（Hubert）、戈利（Golledge）和科斯坦索（Constanzo，1981）的看法，这表示将一个空间接近性的测量与另外一个

在某些特殊属性[3]上的相似性的测量交叉相乘。令 S_{ij} 为两个观测值 i 和 j 的空间接近性的测量，同时令 U_{ij} 表示所关注的某个潜在变量的相似性。向量内积的统计量的一般形式为：

$$\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}U_{ij}\text{，}\ \forall\, i \neq j$$

如果相似性 U_{ij} 被定义为某潜在变量均值正态化后得到的内积(mean-normalized cross-product)，比如 $[(y_i - \bar{y})(y_j - \bar{y})]$，那么，经过适当的比例调整，再将所有观测值的这个量加总，就可以得到一个空间相关性的测量，称之为莫兰(Moran)$\mathcal{I}$统计量。如果 U_{ij} 被定义为差值的平方，比如 $(y_i - y_j)^{[2]}$，得到的统计量就被称为吉尔里(Geary's)$\mathcal{C}$。在本书中我们主要关注莫兰(Moran)$\mathcal{I}$。[4]

例如，在测量民主的例子中，空间相关性将涉及测量国家之间一些空间指标上的邻近程度(比如国家之间是否在200 千米内接壤)和每对国家在民主得分上的相似性。这些统计指标将有助于发现或者探索空间模式。可能它们最有用的地方在于探索诊断那些原本模型中没有考虑到的空间模式的残差。

估算这些相关性的首要目标是确定数据之间的相互关联作用。这要求我们给出观测值之间如何关联的列表。尽管这一步非常重要，但除此处之外，我们将不会进行过多的阐述。数据之间的连接可以通过物理上的距离来建立，比如首都之间的距离。然而，其他的传输途径，比如公路、铁路、水路以及航运等交通网络的密度在某些情况下可能是一个更好的连接指标。类似的，除了首都城市之间的距离，学者们也曾使用两个相邻国家之间的边界长度，作为它们交往机

会的测量。在格里蒂奇和沃德(Gleditsch & Ward, 2001)的研究中开发了一个数据库来记录世界上所有国家之间的最短距离。这里我们将使用该数据,如果国家之间的最小距离小于200公里,那么认为它们是邻国。[5]

表1.3按照两种方式列出了这些数据的一个子集,首先按照列表形式,然后按照矩阵形式。很多电脑程序将大的矩阵按照列表的形式排列,因为这样仅仅将非零的元素保存在记忆体中,可以更有效地存储信息。实际上,对于小的子集,列表存储方式更便于存储数据,且更有利于推导空间特征。

表1.3 欧洲国家子集的连接矩阵

条列清单式

国 家	连接线
丹 麦	德国、挪威、瑞典
芬 兰	挪威、瑞典
法 国	德国、意大利、英国
德 国	丹麦、法国、意大利、瑞典
意大利	法国、德国
挪 威	丹麦、芬兰、瑞典
瑞 典	丹麦、芬兰、德国、挪威
英 国	法国

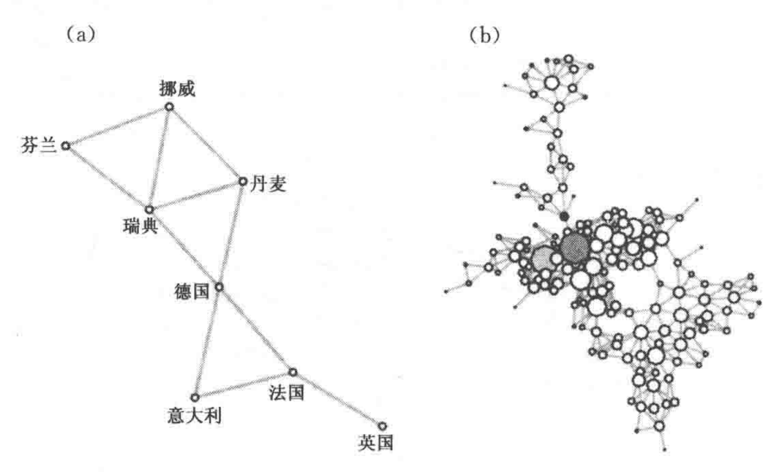
连接矩阵格式

	丹麦	芬兰	法国	德国	意大利	挪威	瑞典	英国
丹 麦	0	0	0	1	0	1	1	0
芬 兰	0	0	0	0	0	1	1	0
法 国	0	0	0	1	1	0	0	1
德 国	1	0	1	0	1	0	1	0
意大利	0	0	1	1	0	0	0	0
挪 威	1	1	0	0	0	0	1	0
瑞 典	1	1	0	1	0	1	0	0
英 国	0	0	1	0	0	0	0	0

然而,每个表都能很容易地被转换成一个方形矩阵,该矩阵描绘了沿行和列排列的观测值以及矩阵内部的关联。矩阵

形式同样有利于明确变量定义或反映空间结构和变化的测量。表 1.3 的第一部分表示一组列表形式的连接数据，第二部分是相应表示二元连接的矩阵 C。

这些数据也可以通过简单网络图来展示，如图 1.5。这种图非常具有启发性，但是一旦节点数量很多，它们就会变得难以阅读。图 1.5(b)就展示了 158 个国家的网络图的拥挤情况。然而，这种视觉网络表现法可能会对某些数据，尤其是小数据的情况非常有用。



注：(b)图中美国是黑色而俄罗斯为灰色阴影。节点大小与 200 千米以内国家的数量成比例。

图 1.5 表 1.3 中 158 个国家数据的简单网络：
(a) 8 个欧洲国家之间的关联；
(b) 158 个国家之间的关联

一旦构造一个观测值之间连接的可能网络，用列表 L 表示或者用连接矩阵 C 表示，我们就可以试图找出所关注的某个变量的取值（这里用 y 来表示）是否与相互连接或者相邻的观测值相似。一种可行的方法是观察两个相连的观测值 i 和 j 是否相似，比如判断 i 观测值的高低是否和 j 观测值的

高低共变。通常 i 会和很多观测值相连，除非它和很多邻近观测值都相似，否则空间聚集效应也不会存在。为了整合相互连接的观测值的有关信息，我们通常假定所有的邻近观测都具有相同权重，并且每一个的权重都将是 1 比上总的连接数量的比例。计算空间滞后(spatial lag)的主要目标，是得出周围区域的均值。美国周围邻国的平均民主得分是多少呢？加纳邻国的平均人均 GDP 是多少呢？这些邻国的均值与每个国家自身的民主得分或者人均 GDP 有关吗？我们提供了一个探索性的统计量，来测量空间相关性。研究者可以以同样的方式生成独立变量之间的相关矩阵，这种空间相关可能会为观测数据提供探索性的信息。

令 y_i^s 表示 y 的所有相关观测值的均值或者平均数，或者叫 y 在空间上的“滞后”(lag)。矩阵表达方式将有助于发现基于 y 所建立的空间滞后 y_i^s 以及连接矩阵 C。我们可以构造一个行标准化的连接权重矩阵 W，该二元连接矩阵 C 中将每一个行向量 c_i 除以总的连接数 $\sum c_i$，使每一行加起来为 1。表 1.4 中给出了一个例子。

表 1.4　含有 8 个欧洲国家的数据子集的行标准化连接矩阵

	丹麦	芬兰	法国	德国	意大利	挪威	瑞典	英国
丹　麦	0	0	0	1/3	0	1/3	1/3	0
芬　兰	0	0	0	0	0	1/2	1/2	0
法　国	0	0	0	1/3	1/3	0	0	1/3
德　国	1/4	0	1/4	0	1/4	0	1/4	0
意大利	0	0	1/2	1/2	0	0	0	0
挪　威	1/3	1/3	0	0	0	0	1/3	0
瑞　典	1/4	1/4	0	1/4	0	1/4	0	0
英　国	0	0	1	0	0	0	0	0

注：连接存在于国境线距离小于 200 千米的国家之间。

在这种情况下，标量 $y_i^s = c_i y$（通过相加）计算了元素 i 所有相邻观测值的平均值或均值，这通常称为空间滞后。$y^s = Wy$ 这组关系告诉我们每一个 y_i^s 和其他国家的 y 以及连接权重 w_i 之间的关系。表 1.5 列出了民主变量的 10 个最大的正向和负向空间滞后。例如，巴林的民主得分为－8，但是它周围的所有邻国都是具有最大负向民主得分－10。另一方面，爱尔兰和葡萄牙具有最高民主得分，而它们的邻国也一样。

表 1.5 最大和最小空间滞后的 10 个国家

国　家	民主	空间滞后	国　家	民主	空间滞后
最大负向空间滞后			最大正向空间滞后		
巴　林	－8	－10	卢森堡	10	9.8
塔吉克斯坦	－1	－7.1	瑞　士	10	9.8
阿　曼	－9	－6.7	英　国	10	9.8
吉尔吉斯斯坦	－3	－6.6	比利时	10	9.8
阿联酋	－8	－6.5	荷　兰	10	9.8
乌兹别克斯坦	－9	－6	加拿大	10	10
卡塔尔	－10	－5.8	斐　济	5	10
也　门	－2	－5.5	法　国	9	10
科威特	－7	－5.3	爱尔兰	10	10
以色列	10	－5	葡萄牙	10	10

注：对应于每个国家，列出的是它们相应的民主得分和空间滞后的民主得分。

第 6 节 | 接近性测量

对很多社会科学家来说，空间分析中最重要的一步是测量个体间的接近性。在社会环境中，距离是什么呢？比如，尽管物理学家们可以用严格的地理或者欧几里得距离测量树和树之间的距离，但测量距离在社会科学分析中却要复杂得多。例如：美国和墨西哥之间隔多远呢？如果我们使用严格的连续性测量，这两个国家是最理想化情况下的邻国，因为它们具有共同的边境线。但是加拿大和美国也具有共同的边境线。这表明它和美国的距离与前者一样大吗？从华盛顿特区到墨西哥城的直线距离是 3000 千米，但是从华盛顿特区到渥太华的距离为 700 千米。我们也可以使用国家之间边境线的长短或者最大的 10 个人口中心的平均距离来测量国家之间的距离。图 1.6 给出了这两种不同识别方法的区别。一些国家版图的中心(空心圆点)和它们的首都(黑色圆点)之间的距离非常远，但是，在一些小国就不会出现这种问题。中国、加拿大、俄罗斯、澳大利亚和美国都属于两种圆点之间距离较远的类型。相对而言，在朝鲜和韩国，版图中心和首都之间的距离就非常小。

在实际应用中，另外一个重要的问题是如何处理缺失的空间数据。插补方法(imputation)就是一种处理方法，当然

也有其他办法(Griffith，2003)。真正的问题在于社会科学的数据通常存在缺失问题,而且很少是随机缺失的。在不是空间数据的情况下,这种缺失可以通过标准方法处理——利用插补方法,或者更常用的是将删除有缺失信息的观测值。然而在空间数据结构中,这些缺失可能会在空间图上产生一些“洞”(holes),从而使我们不能准确完整地表现空间上的接近性。在空间结构上,另一些可能出现的问题是,某些观测值不与其他观测值相连。例如,新西兰在200千米之内就没有其他独立的国家。有两种避免出现这种情况的常用的办法。通常我们会从分析中删去岛国,这是因为在很大程度上它们没有连接对象,从而也不会影响研究中其他观测值的空间过程。更显而易见的原因是,删除它们之后,可以排除空间加权矩阵中的奇异阵情况(也就是行或者列完全由0组成)。第二种方法是选取岛屿最邻近或者最可能的邻国,比如将澳大利亚作为新西兰的邻国连接,即便其他观测值都选取200千米,作为它们之间是否有连接的标准。更一般的情况是,可以对所有的个体选取最近的 k 个邻国距离。

图1.6可以通过下列的R命令生成。

```
# Set working directory
dd <- c("C:...")
setwd(dd)

# Plotting map with centroids and capitals

# Load required libraries
library(RColorBrewer); library(maptools)
```

```
library(spdep); library(sp); library(rgdal)

# Read a Robinson projection map from an ESPI shapefile
rob.shp <- read.shape("wg2002worldmap.shp")

# Indicate the id codes for each polygon/country
rob.map <- Map2poly(rob.shp, region.id =
    unique(as.character(rob.shp$att.data$FIPS_CNTRY))))

# Indicate the map projection
tr <- readShapePoly("wg2002worldmap", IDvar = "FIPS_CNTRY",
    proj4string = CRS(" + proj = robin + lon 0 = 0"))

# Extract the relevant variables and exclude missing data
ct <- na.omit(rob.shp$att.data[,c(1, 18:20)]

# Assign relevant variable/column names
colnames(ct) <- c("ID", "x", "y", "City_POP")
ct$x <- as.numeric(as.character(ct$x))
ct$y <- as.numeric(as.character(ct$y))

# Add coordinates
coordinates(ct) <- c("x", "y")
proj4string(ct) <- CRS(" + proj = longlat + datum = WGS84")

# Transform the coordinates to the robinson projection
ct_rb <- spTransform(ct, CRS = CRS(" + proj = robin + lon0 = 0"))

# Replot the map itself without a bounding box
plot(rob.map, border = "Grey", forcefill = T, xaxt = "n", yaxt = "n",
    bty = "n", lwd = .000000000125, las = 1, ylab = "",
```

```
    main = "Centroids and Capitals", xlab = "")

# Add the centroids
points(coordinates(tr), pch = 19, cex = .5, col = "grey")

# Add the capitals
points(coordinates(ct_rb), pch = 19, cex = .5, col = "black")

# Add segments between centroids and capitals
tr_or <- coordinates(tr)
rownames(tr_or) <-
    as.character((attributes(tr)$data)$FIP S_CNTRY)
ct_rb_or <- coordinates(ct_rb)
rownames(ct_rb_or) <- as.character(ct_rb$ID)

# Delete Kiribati (91), as longitude extends across
    international date line
coor_dif <- cbind(tr_or[-91,], ct_rb_or[rownames(tr_or),][-91,1])
x1 <- coor_dif[,1]
x2 <- coor_dif[,3]
y1 <- coor_dif[,2]
y2 <- coor_dif[,4]
segments(x1, y1, x2, y2, col = "slategray4")
```

以上我们列举了两种测量距离的基本方法，但是这还停留在表面。对距离的度量，也可以用平均行程时间、每两点之间的移动电话通话数目、两地之间的旅游观光数量，或者其他任何形式的关于距离与交往的测量。例如，国家之间有大量的贸易活动，就可以被认为在经济上“密切”(Lofdahl,

2002)。格里菲斯(Griffith, 1996)就提供了一些这种测量的想法和操作方法。



图 1.6 地理中心(空心点)和首都(黑点)之间连线图

我们通过 y 和 y^s 之间的相关程度,来测量国家自身的民主发展水平和它们邻国的民主水平的关联,这看起来是一件合乎常理的事情。某国自身取值与各邻国的加权平均值之间的线性相关被称为莫兰 I 统计量(Moran, 1950s, 1950b),这个全局相关(global correlation)包括一个观测值和它邻近值之间的所有取值。广义的莫兰 I 统计量用一个加权的成比例的向量内积表示:

$$I=\frac{n\sum_{i}\sum_{j\neq i}w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{(\sum_{i}\sum_{j\neq i}w_{ij})\sum_{i}(y_i-\bar{y})^2}$$

其中 w 表示行标准化的加权矩阵 W, y 是我们所关注的变量。

I 被认为服从正态(渐进)分布,其均值为 $-1/(n-1)$。莫兰 I 的方差可表示为:

$$\operatorname{var}(I)=\frac{n^2(n-1)\frac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})^2-n(n-1)\sum_{k}(\sum_{j}w_{kj}+\sum_{i}w_{ik})^2-2(\sum_{i\neq j}w_{ij})^2}{(n+1)(n-1)^2(\sum_{i\neq j}w_{ij})^2}$$

如果将所关注的变量标准化为 z_i，莫兰 I 可以简化为：

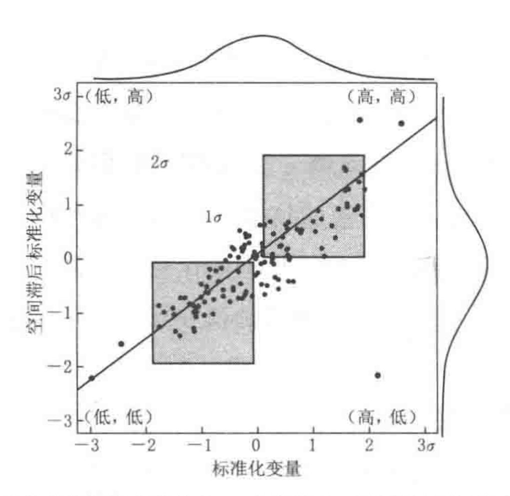
$$I=\frac{1}{2}\sum_{ij}c_{ij}z_iz_j \quad \forall i\neq j$$

莫兰 I 统计量通常通过建立包含均值和方差信息的 z 值来检验空间相关性。

莫兰 I 并没有一个真正固定的度量标准，但是它的期望值是 $-1/(n-1)$，而不是 0。然而，莫兰 I 统计量可以用图形来解释，帮助我们理解个体之间的不同空间相关程度如何导致该统计量的取值变化。比如一个 $\tilde{y}$ 和其邻近值平均数 $\tilde{y}^s$ 的散点图[这里我们使用标准化的 $\tilde{y}=(y-\bar{y})/sd(y)$，使得均值为 0、标准差为 1]。在这个图中，观测值在四个象限里围绕 $\tilde{y}$ 和 $\tilde{y}^s$ 的均值分布，这反映了变量 y 的空间相关关系。如果 y 没有空间聚集关系或者存在相关关系，y^s 的值将不会随着 y 的变化而发生系统性的变化。然而，如果存在正向空间相关，个体观测值的取值高于或者低于 y 的均值将在 y^s 上(或者说在邻近国家中)相应地反映了高或低[①]的趋势。大部分的点将落在第一和第三两个象限，在这些位置，个体和它们的邻近值是相似的，而位于第二或第四象限的观测值将会比较少。如果我们对散点图画出它们相应的回归直线，它的斜率就是针对原始变量 y 和连接列表或者矩阵 C 之间的莫兰 I 相关系数。

图 1.7 以一种固定格式说明了莫兰 I 统计量，并解释了变量及其一阶空间滞后量组成的散点图。回归直线的斜率表示数据中空间相关性的平均数；也就是莫兰 I 统计量。

① 原书为低或高，应该为笔误。——译者注



注：变量经过标准化使得其均值为 0，方差为 1。同质性的邻近观测值聚集群用阴影部分标示。图中也给出 OLS 回归直线。

图 1.7 变量和其空间滞后的散点图

莫兰 I 通过调整 y 的变化以及每个观测值相邻点的数量，比较了 i 的所有邻近点和均值之间偏差的关系。莫兰 I 值越高，表明地理上的聚集作用越强；也就是说，邻近取值的相似性越大。这个统计量测量了一个观测值和它的邻近点之间的平均相关关系。图 1.7 就解释了这个基本概念。空间滞后情况（某点的邻近点的平均值）用纵轴表示，而横轴表示每个观测值的取值，标准化以后使得其均值为 0、方差为 1。方框表示±2，大部分的观测值都落在这个边界内（注意：在这里 $2\sigma = 2$，因为变量经过了标准化）。那些落入阴影方框的观测值表明其拥有同质性很强的邻国。那些落在高于（0，0）和（2，2）阴影区域的观测值表明其取值高于均值，并且它们的邻国的平均取值也高于均值。类似的，落在（0，0）和（−2，−2）[①]之间阴影区域内的观测值低于均值，同

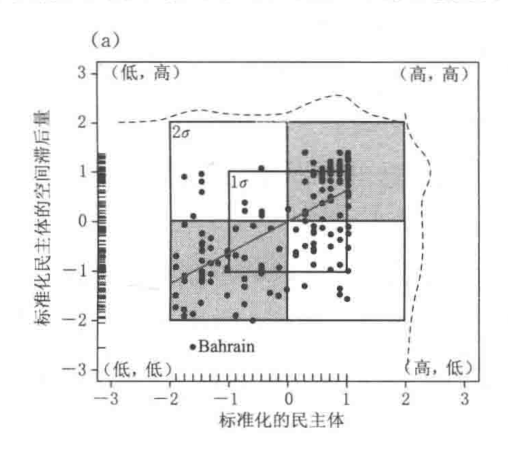
① 原书为（−2，2），根据上下文应该是笔误。——译者注

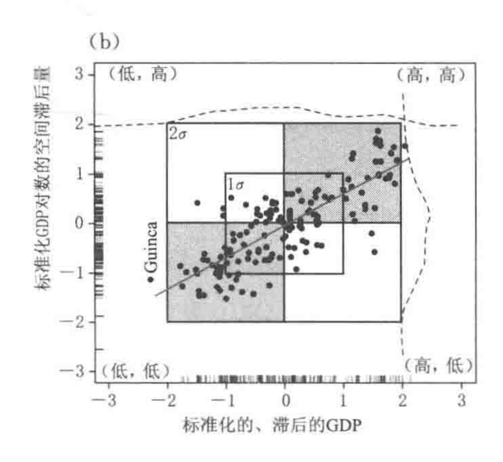
时它们的邻近点也如此。这个散点图的对角线区域有很多观测值，这反映了相似值的聚集效应。图中也有一些点所代表的观测值自身低于观测变量的平均值，但是，平均而言，其邻近的点却远高于变量的均值。在散点图的左上角部分的一个点就是这种情况，它可以被看作是一个被取值很高的邻近点包围的、而自身取值很低的孤立点。针对这些标准化后的观测值的OLS回归，可以生成一个概括观测值及其邻近点之间取值关系的测量值。举个例子，如果我们关注的变量为犯罪率，并收集了休斯敦市每个管辖区的数据，在图的右上部分将是具有高犯罪率的管辖区，其周围也是高犯罪率的地区。类似的，左下角代表自身犯罪率低而周围犯罪率也低的地区。穿过这些标准化点的回归直线的斜率就是莫兰I统计量。

针对莫兰I系数的统计检验，需要加入其他假设，因为在简单概率检验的经典框架中，需要用到一阶和二阶矩（均值和方差）。在假设检验框架中，经常被忽略的重要一步，是明确零假设。对空间模型而言，由于在不同的空间变量中可能存在各种迥异的模式，因此零假设就显得很不明显了。比如，空间模式是正态分布的吗？还是随机分布的？如果是随机分布的，那么在空间中是完全随机的吗？一般来说，在当前文献中，有两种解决办法，但它们在一定程度上都是事先设定的（ad hoc）。第一种方法假设数据是正态分布的。克里夫和奥德（Cliff & Ord, 1971）计算出这种情况下I的方差。尽管很多研究表明假设莫兰I为正态分布通常是错误的（Boots & Tiefelsdorf, 2000; Tiefelsdorf, 1972），但是大多数软件和应用文章仍然使用正态性假设[6]。第二种方法是利

用蒙特卡洛模拟(Monte Carlo simulation),对连接矩阵随机出足够多次的行列变换,从而得到一个随机化的零模型。大多数统计软件都会提供这两种主要方法供选择,它们通常(当然也不总是)会得到相似的结果。

图 1.8(a)给出了民主得分标准化后的散点图。该图将 $\pm 2\sigma$ 用方框标出,为了便于发现哪些观测值过度异常。右上





注:(a)民主化,莫兰 I = 0.64;(b)人均 GDP,莫兰 I = 0.65。

图 1.8 标准化变量与其空间滞后图

象限表示得分较高并且周围也是较高得分的观测值。位于左下象限的点,表示得分较低并且周围也是相似较低得分的个案,巴林就是其中一个极端的例子。在图 1.8 中,非对角线上的点表示这些国家和它们邻国的民主水平差别很大。正如图中所示,对于专制国家这样的个案是非常少的(最例外的情况是白俄罗斯),而对于民主国家这样的个案更是少之又少。图中也画出了回归直线,它的斜率就表示民主的莫兰 I 统计量(为 0.64),这个值比该统计量在这个例子中的期望值要大得多(−1/158)。图 1.8(b)表示人均 GDP。几内亚位于左下角,已经超出了 2σ 的方框范围;卢森堡在右上角也超出了该方框区域。

我们可以根据表 1.2 中 OLS 估计出的残差的莫兰 I 来考察残差的变化是否表现出空间聚集[7]。这仅仅是一个探索性的考察,在原始数据中使用莫兰 I 的作用差不多。利用 R 软件可以很容易地完成,定义一个回归目标 ols1.fit 并取列表 nblist 中 200 千米边界范围内的其他国家作为其"邻国"。

```
source("chapter1data.R")
ols1.fit <- glm(democracy ~ log(gdp.2002/population), data = sldv)
library(spdep)          # Load spdep library for moran.test()
moran.test(resid(ols1.fit),nb2listw(nblist))
lm.morantest(ols1.fit,nb2listw(nblist))
```

根据 OLS 残差计算出的莫兰 I 统计量为 0.40,方差为 0.0028。相应的标准分为 7.77,它比 −1/158 要大得多,同

时对应的 p 值≈0。这表明基于观测值之间相互独立的假设得到的 OLS 结果受到因变量和自变量的空间聚类的强烈影响。因此，这可能误导了我们在统计上和实际上推断民主和社会财富(由人均 GDP 表示)之间的关系。

第 7 节 | 估计空间模型

空间分析的一系列步骤简单来说是怎样的呢？

第一，将数据在地图上标示出来，尤其是因变量。这可以在很多种环境中完成，比如数据表插件（spreadsheet plugin）、地图混搭程序（map mashups），以及 GIS 软件包，但是最好的情况应当是能够统计分析的平台。我们将介绍 R 库的使用，尤其是通过 maptools 和 spdep 创建变量分布的简单地图。

第二，同时，判断因变量上是否有明显的空间相关。对于本书中的大多数方法[即不是点过程（point processes）的方法]，这是指计算莫兰 I 统计量来估计空间相关的大小。分析者在有的情况下希望通过局部空间相关指标（Local Indicator of Spatial Association, LISA）来考察和绘出每个观测点对空间相关的贡献。本书将不详细讨论这一点。更多有关的讨论和例子，可参见格里蒂奇和沃德（Gleditsch & Ward, 2000）、安瑟林（Anselin, 1995），以及奥德和格蒂斯（Ord & Getis, 1995）的文章。

第三，将这些空间滞后变量准确地合并到基本的统计框架中，并且检验得到的残差是否仍然是空间相关。

第四，除了利用正态模型探索方法估计模型拟合程度和预估参数的不确定性程度，我们也需要计算和检验均衡效应

(equilibrium impact)。这表明需要梳理出估计得到的空间模型的均衡效应和相互反馈的作用。

下面我们将根据当前的例子来讲解这些步骤。

将数据地图化同时建立空间权重矩阵

我们已经讲解了如何将 2002 年 158 个国家民主得分的数据在地图上标示出来。在这一节当中,我们将讲解如何将 OLS 回归残差在地图上标示出来。在图 1.3 和图 1.4 中,数据自身在地图上就已经标示出来了。我们同时发现,使用收入对民主进行回归之后,其残差也表现出空间相关性。我们通过定义 200 千米距离内的“邻国”计算得到莫兰 I。如前面提到,该例子中莫兰 I 为 0.4,其方差为 0.0028。在一般情况下这种显著结果让我们确信从图 1.3 和图 1.4 中观测到的空间模式实质上影响了回归结果,也就是说,它造成了估计和标准误的偏差。图 1.9 反映了 OLS 中的残差。

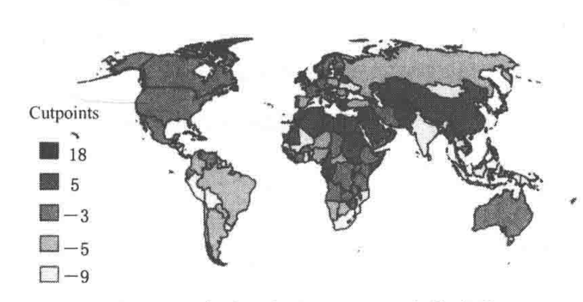


图 1.9 在地理上显示 OLS 回归的残差

寻找空间类型

根据锡恩(Shin, 2001)的研究,我们也可以创建一个锡

恩空间散点图(Shin spatial scatterplot)。该图画出了每个输入变量标准化后的值——在这里为残差——和相对应的空间滞后或者是与其相连接的观测值的均值(图 1.10)。阴影框表示当某值大于残差均值时,其邻近点也为正值的情况。坐标轴上包括一个"毯图"(rug plot)用于表明变量的分布。外部边缘区域表示对变量本身以及其空间上估计分布的核密度分布(kernel density)。生成此图的代码如下:

```
pdffilename <- c("file name and path")
pdf(file = pdffilename, width = 5.0, height = 5.0, family = "Times")
dem <- (resid(ols1, fit))  # residuals
ds <- (dem - mean(dem))/sqrt(var(dem))  # standardized democra-
        cy score
# create spatial lag and standardize it
ds.slag <- as.vector(wmat % * % ds)
ds.slag <- (ds.slag~mean(ds.slag))/sqrt(var(ds.slag))
plot(ds, ds.slag, xlim = c( - 3, 3), ylim = c( - 3, 3), pch = 20, las = 1,
            xlab = "standardized democracy",
            ylab = "spatial lag of standardized democracy")
reg1 <- lm(ds.slag~ds)

# establish a grid
xgrid <- sq( - 3, 1.5, length.out = 158)
x0 <- list(ds = xgrid)
pred.out<- predict(reg1, x0, interval = "confidence")

# put 1 and 2 sigma boxes on plot
lines ( c( - 2, - 2, + 2, + 2, - 2), c( - 2, + 2, + 2, - 2, - 2))
```

```
lines ( c( -1, -1, +1, +1, -1), c( -1, +1, +1, -1, -1))
lines ( c( -2, +2), c(0, 0))
lines ( c(0, 0), c( -2, +2))

# some text for context
text( -2.5, 3, "(low, high)"; text(2.5, 3, "(high, high)")
text( -2.5, -3, "(low, low)"; text(2.5, -3, "(high, low)")
polygon(x = c( -1, 0, 0, -1), y = c( -1, -1, 0, 0), col = "slategray3")
polygon(x = c(0, 1, 1, 0), y = c(0, 0, 1, 1), col = "slategray3")

# plot c. i. region
polygon(x = c(xgrid, rev(xgrid)), y = c(pred. out[, 3],
          rev(pred. out[, 2])), col = "slategray3", border = T)

# put data on plot
points(ds, ds. slag, pch = 20)

# densities
sldensity <- density(ds. slag)
lines(sldensity $ y + 2, sldensity $ x, lty = 2, col = "slategray4")
ddensity <- density(ds)
lines(ddensity $ x), ddensity
$ y + 2, lty = 2, col = "slategray4", xlim = c( -2, 2))
points(ds, ds. slag, pch = 20)
lines(xgrid, pred. out[, 1], type = "1", lty = 2, col = "gray80", lwd = 2)

# rugs on two sides
rug(jitter(ds, factor = 2), col = "slategray3")
rug(ds. slag, side = 2, col = "slategray3")

# label some points
```

```
text(-2., -2.3, "Oil Exporters", col = "slategray4")
dev.off()
```

接下来我们将考察民主测量上的空间关联性。民主变量的空间滞后，简单来说，就是周围国家的民主水平的平均数。这里邻国民主得分高的国家，其自身的分值也高，同时其邻国专制程度高的国家，其分值也负得越大。我们在图 1.11 中将它们在地图中画出来。该地图显示，位于非洲和亚洲的国家，其邻国都是非民主国家，而欧洲和大部分美洲国家的邻国都为民主国家(Gleditsch, 2002a)。

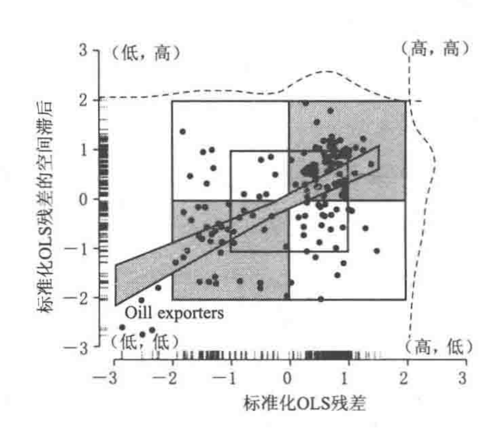


图 1.10　OLS 残差锡恩图(Shin Plot)

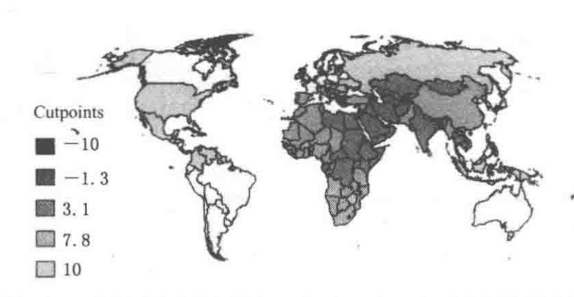


图 1.11　民主的空间滞后性(颜色越深表示空间滞后变量值越大)

其执行代码如下：

```
# sldv2 is the data. frame
# mdd2 is the minimym distance data. frame
nblist <- vector(mode = "list", length = dim(sldv2)[1])
attr(nblist, "region. id") <- sldv2 $ tla
attr(nblist, "class") <- "nb"
nbnms <- data. frame(sldv2 $ tla, c(1:dim(sldv2)[1]))
names(nbnms) <- c ("acr", "nm")

min200 <- mdd2[mdd2 $ mindist< = 200,] # Create an index of
          the isolates
nodata <- setdiff(sldv2 $ tla, unique(c(min200 $ ida, min200 $ idb)))

# Find neighbors for each row in the sldv for(i in 1:dim(sldv2)[1]){
  temp <- min200[min200 $ ida = = sldv2 $ tla[i] |
        min200 $ idb = = sldv2 $ tla[i],]
cty <- unique(c(temp $ ida, temp $ idb))
cty <- setdiff(cty, sldv2 $ tla[i])
  nblist[[i]] <- nbnms[match(cty, nbnms $ acr), "nm"]
}

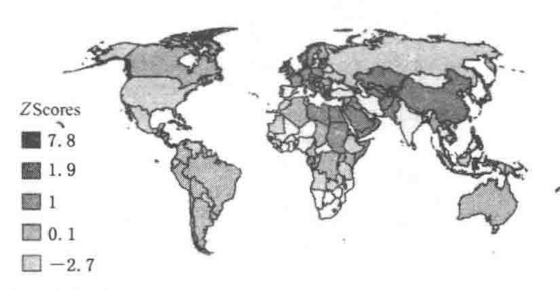
# wmat is the row standardized weights matrix
wmat <- matrix(0, ncol = dim(sldv2)[1], nrow = dim(sldv2)[1])
rownames(wmat) <- colnames(wmat) <- sldv2 $ tla
for (i in 1:dim(min200)[1]){
  wmat[min200 $ ida[i], min200 $ idb[i]] <- 1
}
```

```
wmat <- wmat/rowSums(wmat)

# Calculate the spatial lag of democracy
Democracy.spatial.lag <- as.vector(wmat %*% sldv2$democracy)
```

除了在地图上标示出一阶民主空间滞后，将每个观测值对总体莫兰 I 统计量的贡献在地图上标示出来也是很有用的。这个量就叫做 LISA(局部空间相关指标)统计量。这里我们将所有的值标准化，然后将它们在地图中标示出来，如图 1.12。局部莫兰统计量的计算来自于奥德和格蒂斯(Ord & Getis, 1995)、安瑟林(Anselin, 1995)以及格蒂斯和奥德(Getis & Ord, 1996)的文章。

该地图表明，从其邻国的民主发展水平来看，哪些国家的情况比较异常。南非和西非就属于这种情况，印度也是。



注：用灰色等级指代邻国之间相似性。

图 1.12 局部莫兰 I 统计量的 z 值

第 8 节 | 小结

首先经过仔细考察数据并将数据视觉化以后，我们基于民主水平和财富（用人均 GDP 的对数来表示）之间关系的线性假设，得到 OLS 回归结果。根据从回归残差中得到的有力证据，我们发现，残差之间表现出空间聚集的关系，这违背了回归中认为个体观测值的误差项应当相互独立的假设。OLS 估计基于个体相互独立的假定，因此，在分析收入和民主之间的关系时，OLS 将不会是一个可靠的模型。更重要的是，模型中假定对于这些独立观测值，民主仅仅受到收入的影响，而忽略了显而易见的地理聚集效应的特征。我们也展示了如何利用地图和简单的统计量来提供有关空间聚集程度及其特性的探索性信息。

即便研究者对回归分析不感兴趣，也可以用社会科学数据发现空间模式。如果仅仅进行简单的均值检验或者用回归方法分析空间排列的数据，而不考虑空间相关关系，都将会带来错误的推论，从而导致错误地拒绝了已有的假设。

通过地图展示相关联的数据，为我们提供了一种判断空间模式是否存在的探索性方法，而空间模式的存在，也会使统计推论变得更加复杂。下面我们将介绍如何估计包含空间滞后因变量的回归模型，这种方法有助于我们明确地将空间依赖关系纳入回归框架中。



第2章

空间滞后因变量

在本章中，我们将讨论如何将空间依赖性，也就是将包含“空间滞后”因变量的 y 明确地加入回归方程的右边。这种模型有很多不同的名称。安瑟林(Anselin, 1998)将其称之为空间自回归(spatial autoregressive)模型，但这个术语可能会引起一些混淆，因为自回归这个词在地理统计文献中表示另外一些完全不同的空间模型。为了简单起见，这里我们称之为空间滞后 y 模型，原因是该模型的主要特征是加入了作为协变量的空间滞后因变量。

当我们确信每个个体 i 的 y 值受到其“周围值”的直接影响的时候，空间滞后 y 模型就很适用。这种影响大于和超过 i 的其他协变量的影响。如果我们相信 y 并没有直接受到周围值的影响，而是因为某些在模型识别中忽略的空间聚集特征同时影响了个体 i 的 y 值及其周围值，这时就需要考虑一种空间相关误差的模型，我们之后会讨论到。在空间滞后 y 模型下，因变量 y 必须被视为连续变量。本书将不讨论更复杂的二分因变量的例子。这是因为更复杂的情况可能无法得到闭合形式的解，同时迭代估计方法也超出了本书的范围(参见 Ward & Gleditsch, 2002)。

第1节 | 空间滞后因变量的回归

为了进一步讲解空间滞后 y 模型，我们回到前面世界范围内民主分布的例子。我们已经看到民主的分布展现出空间聚集的效应，也就是如果一个国家周围都是民主水平很高的国家，那么它自身的民主 POLITY 得分也会较高。尽管有一些民主聚集现象可能来自于人均 GDP 的空间聚集，并且它和民主之间也具有正向相关，但是我们仍然发现：即便是控制了国家的人均 GDP 水平之后，民主的空间聚集效应依旧没有完全消失。在民主对人均 GDP 回归模型中，由于假设误差 ϵ_i 相互独立，这样利用回归残差就可以检验空间依赖是否存在；也就是说根据 $\hat{\epsilon}_i=(\hat{y}_i-y)$，同时利用莫兰 I 相关系数以及识别出的矩阵 C 的连接形式——在这里国家之间如果在 200 千米范围以内，就看做相互连接。在本例中，我们发现了残差间的强空间相关关系。残差的莫兰 I 统计量为 0.40，相应的 z 值近似于 8[8]。这个结果远远大于零假设(即空间独立性)为真的情况下的值。换句话说，这表明国家民主发展水平及其地理上邻国之间的正向相关关系，远远大于我们所预期的、人均 GDP 的解释作用。这种结果是相当典型的，仅仅靠空间聚集的协变量，并不能完全去除研究中因变量的空间聚集效应。

假设在控制了国家的人均 GDP 以后，民主的分布依然表现出空间聚集效应，我们就需要通过可能的方法将空间依赖性纳入回归模型。与时间上的序列聚集情况一样，我们可以将空间自相关看做外在干扰或本质现象。空间依赖性将使得对人均 GDP 的 $\hat{\beta}$ 及其标准误的估计出现错误，这是由于在相互关联的个体之间，误差不能被视为相互独立。原则上，在用人均 GDP 对民主进行估计的例子中，这些问题可以通过某些考虑到空间误差相关性的估计值来解决；也就是说，残差的变化并不能完全被人均 GDP 解释。这种方法通常被称为空间误差模型，我们下面将会谈到这种方法。

然而，这里更大的关注点是什么影响了民主，而不仅仅是估计国家人均 GDP 与其民主发展之间的相关作用。如果一个国家的民主水平看上去与其邻国的民主水平相关，这就为我们提供了有关民主本身分布的信息，同时还让我们有机会了解空间依赖关系对民主的促进和阻碍作用。正因为如此，一个更可能和有趣的方法是将空间相关看作民主的本质特征，而不是一种统计干扰。

这里观测到的空间相关，表明观测值之间存在空间依赖，因此，国家 i 的民主值由于其周围国家 j 的民主程度不同而存在很大差异。与其将国家 i 的民主看做仅仅受人均 GDP 的影响，不如设计一个模型，将民主视为其自身人均 GDP 和周围国家民主水平的函数，定义为 $w_i . y_i$，这里对于所有同 i 相连接的国家 j，表示连接性的向量 w_i。比如，矩阵 W 中的行 i 中的值都必须为非零值。前面我们曾提到，在表示连接的矩阵 W 中行经过了标准化，使得每一行的值加总为 1。

这种推导表明空间滞后因变量的模型具有如下形式：

$$y_i = \beta_0 + \beta_1 x_i + \rho w_i \cdot y_i + \epsilon_i \qquad [2.1]$$

这里，空间滞后(ρ)的参数如果为正值，表明国家应该有更高的民主值，如果它们周围的国家的民主平均得分也很高。

这让我们容易把空间滞后 y 模型联想为类似的时间序列自回归模型，其中时间序列相关性的表示是通过在方程右边加入一个滞后的因变量 y_{t-1} 来估计其他协变量(比如 x_t)对 y_t 的影响。在空间滞后 y 模型中，系数 $\hat{\beta}_1$ 不同于 OLS 的回归系数，这是因为人均 GDP 对民主水平的估计作用大小控制了 y 的空间依赖关系，或者说国家 i 的民主水平的变化可以通过其他国家 j 的 y 值来解释。因此，在估计 x 变化带来的影响时，我们还需要考虑到空间上的相互影响。

表 2.1 和表 2.2 提供了考虑和不考虑 y 的空间滞后效应后，2002 年 158 个国家民主水平对人均 GDP 自然对数的 OLS 回归。在忽略空间滞后 y 的情况下，我们观察到：OLS 结果中人均 GDP 有非常大的正系数，为 1.68。相比而言，在空间滞后 y 模型中，人均 GDP 对数的系数为 0.76，比原来值的一半还要小，不过按照传统的显著检验标准，它还是远远大于 0。

表 2.1　没有空间滞后 OLS

OLS	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距	−9.69	2.43	−3.99
人均 GDP 对数	1.68	0.31	5.36
$N = 158$			
Log likelihood ($df = 3$) $= -513.62$			
$F = 28.77$ ($df_1 = 1$, $df_2 = 156$)			

表 2.2 具有空间滞后的 OLS

	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距	−4.98	2.07	−2.40
人均 GDP 对数	0.76	0.28	2.72
ρ	0.76	0.088	8.65
$N = 158$			
Log likelihood ($df = 4$) $= -482.48$			
$F = 58.64$ ($df_1 = 2$, $df_2 = 155$)			

估计得到的空间滞后 y 不仅是较大正值(0.76)，而且按照标准它是高度统计显著的。这支持了我们的推测，即一个国家的民主水平和它的地理邻国的民主水平之间存在共变关系。实质上，该模型表示：如果一个国家的邻国的平均民主水平处于最小可能值（比如，−10），与邻国平均民主水平为 0(这接近于 1945 年以来 POLITY 的历史平均得分)的国家相比，该国家的预期民主水平将少 7.6 分。相反，如果一个国家的邻国平均民主得分为 10，与一个邻国平均民主得分为 0 的国家相比，其预期的民主水平将高 7.6 分。这些估计结果反映了我们前面提到的民主的聚集效应。尽管大多数民主国家更可能具有高的人均 GDP，我们也观察到：2002 年，在一些民主聚集的地区，其人均 GDP 并不是很高，比如拉丁美洲，此外也有一些专制国家聚集区拥有较高的平均 GDP，例如波斯湾各国。

与表 2.1 中假设独立观测值的模型拟合度相比，表 2.2 中包括空间滞后 y 的模型对数据的拟合要好得多。该模型比假设观测值彼此独立的模型具有更高的 F 统计值和对数似然值。反过来，这让我们更加相信空间滞后 y 在识别民主分布中起到了重要作用，也就是说，不仅仅是国家人均 GDP

在起作用。然而，探索模型本身并不能令人完全相信空间方法的适用性。空间方法的优越性并不是因为仅仅起到探索性的作用，而是因为它为观测值之间的关联和相互反馈作用建立了一种看似合理的形式。

标准最小二乘法回归具有如下形式：

$$y_i = x_i\beta + \varepsilon_i$$

如果 ε_i 被分解为因变量的空间滞后项——它与因变量相关——和一个自变量的误差项，$\varepsilon_i = \rho w_{i.} y_i + \epsilon_i$，这就是空间滞后模型的形式：

$$y_i = x_i\beta + \rho w_{i.} y_i + \epsilon_i$$

然而，如果换一种方法令 $\varepsilon_i = \lambda w_{i.} \xi_i + \epsilon_i$，得到：

$$y_i = x_i\beta + \lambda w_{i.} \xi_i + \epsilon_i$$

这就是空间误差的形式。

下面我们将介绍空间滞后因变量模型；空间误差模型将留到第 3 章[9]讨论。

我们很容易将表 2.2 中的空间滞后 y 模型对人均 GDP 的系数估计，与表 2.1直接比较，并得出表 2.1 中人均 GDP 效应更大的结论。然而，这种解释并不正确。由于模型在方程 2.1 中加入了空间滞后之后变为一种自回归形式，因此 x 系数的作用反映了 x_i 对 y_i 的短期效应，而不是像不包含空间滞后 y 的 OLS 回归中 x 系数的净效应。由于 y_i 的值将影响其他国家 y_j 的民主水平，同时，这些 y_j 反过来影响 y_i，我们需要考虑到额外的效应，也就是 x_i 通过对其他国家民主水平的影响所导致的对 y_i 的短期影响。

这种解释类似于时间序列模型中协变量 x_t 的系数 β，此时方程的右边包括因变量的时间滞后变量 y_{t-1}，如下：

$$y_t = \beta x_t + \phi y_{t-1} + \epsilon_t$$

在这个方程中，β 代表 x_t 对 y_t 的即时效应，但是反过来，它又会在下一个时间段上影响 y_{t-1}，同时 x_t 的长期效应也必须考虑净效应部分，该效应来自于自回归部分，或者说来自于滞后 y_{t-1} 的估计参数的影响。x_t 的长期效应为 $\beta/(1-\phi)$。当 ϕ 很大时[①]，长期效应 $\beta/(1-\phi)$ 将会显著大于 β。

沿用上面的类比，如果某一个国家 i 的人均 GDP 增加一个单位，这将对该国的民主水平产生直接影响 β_1。然而，方程 2.1 的模型表明：由于国家之间反馈作用产生的空间动态性，国家 i 的民主水平将对其邻国的民主水平产生影响。因此，i 的民主增长将影响邻国 j 的民主水平。同时，反过来，邻国的邻国也将受到影响，并且影响将扩大到所有相互连接的国家。一般来说，所有国家都会拥有一些邻国，因此最终所有的国家都会受到影响。但是注意方程 2.1 中包含了系统 y 中所有国家的民主水平，因此如果连接 i 的其他国家的民主水平提高，i 的民主水平也会提高。假设在一种实验状况下，某观测值受到一个外生的冲击，这种影响将通过观测值之间的相互作用引发一系列调整，并通过在系统中的循环产生回荡效应，直到生成新的稳定均衡（Cressie，1993；Lin、Wu & Lee，2006）。

除了关注空间滞后 y 模型中 x_i 的估计系数，考虑均衡效应也很重要。可惜的是，空间滞后 y 的长期效应并不能像在

① 准确地说，应该是 ϕ 接近于 1 的时候。——译者注

时间滞后 y 情况中那样用简单的形式表示出来。我们会在后文讨论如何描述和估计空间滞后 y 模型中协变量的均衡效应。首先我们将讨论由于空间滞后 y 在方程右边所带来的内生性问题,以及与之相关的最小二乘法步骤中模型估计一致性的问题。

下面一节将利用矩阵代数讨论估计问题,以及为什么最大似然估计量(Maximum Likelihood Estimator, MLE)适合用于估计空间滞后 y 模型。因为使用 MLE 本身并不需要了解本节中的所有内容,读者若对估计问题不感兴趣,可以跳过本节内容,直接进入下一节。

第 2 节 | 估计空间滞后 *y* 模型

在包含时间滞后项 y_{t-1} 的时间序列模型中，如果回归模型的残差不存在序列相关，时间滞后项 y_{t-1} 并不影响 OLS 的估计结果。更准确地说，假设模型被正确识别，包含滞后因变量的 OLS 模型并不导致估计问题。尽管在滞后因变量的优劣问题上已经存在大量争议，但这种争论取决于数据产生过程中的其他特定假设是否也合理。这里可参考基尔和凯利(Keele & Kelly, 2006)的讨论。尽管 y_{t-1} 在时间上早于 t，y 在空间上的滞后却是同时存在的，并且来自于 y 自身。这种共时性(simultaneity)特征导致空间滞后 y 模型估计上的问题。为了理解这一点，我们可以借助矩阵代数形式的空间滞后模型。根据安瑟林(Anselin, 1988)的表示法，空间滞后 y 模型可以写成：

$$Y = \rho \mathrm{W}y + \mathrm{X}\beta + \epsilon$$

$$\epsilon \sim N(0,\ \sigma^2 I)$$

这里 I 代表单位矩阵($n \times n$ 矩阵中，对角线上为 1，其他地方为 0)，$\epsilon \sim N(0,\ \sigma^2 \mathrm{I})$ 表示误差服从正态分布，并且方差一致，即误差与协方差矩阵的积为 0。如果 $\rho = 0$，表明不存在空间依赖，将等式右边的第一项去掉后，我们就得到了标准的 OLS 回归模型。然而，如果 $\rho \neq 0$，由于共时性的存在，OLS

的估计将不会随着样本量的增加而收敛到它们的“真实”值。相反，在OLS形式中被忽略的反馈效应或依赖效应将增大，而不会随着样本量的增大而消失。实际上，这种效应明显依赖于样本量的大小和连接矩阵的具体形式。

如果靠OLS来估计空间滞后y模型存在问题，那么有没有其他估计方法呢？空间滞后y模型可以通过两步工具变量估计法来估计，比如，用外生变量X、WX和W^2X作为空间滞后y的工具变量。在这里我们将不详细介绍如何用工具变量进行估计，而是关注如何在空间滞后y模型中进行最大似然估计。在模型被正确识别的情况下，这可以得到一致的和渐进有效的估计值。尽管由于方程右边Wy的存在，使得空间滞后y的OLS估计存在共时性问题，但是最大似然估计保持了渐进的性质，并且估计的不一致或者偏误大小将随着实践中的具体情况发生变化。弗兰泽兹和海耶斯(Franzese & Hayes, 2007)利用蒙特卡罗模拟方法研究了不同估计值的性质，结果发现：在某些情况下，包含空间滞后y的模型的OLS估计值仍然比最大似然估计具有更小的均方误(mean squared errors)。在小样本情况下，这将很难影响到空间表达式。

最大化空间滞后y模型的似然值是比较复杂的。为了说明这个问题，我们将空间滞后y模型写成如下形式：

$$\epsilon = y - \rho Wy + X\beta = (I - \rho W)y - X\beta$$

反过来，我们可以将估计量β写成：

$$\beta = (X'X)^{-1}X'(I - \rho W)y$$

当ρ未知的时候，要找出该模型的β估计值将是一件困难的

事情，这是因为对数似然函数包含了$|I-\rho W|$行列式。这个ρ的n阶多项式可以通过每一次迭代估算。然而，奥德（Ord，1975）表示如果W存在特征值（ω_1，…，ω_n），那么：

$$|\omega I-\rho W|=\prod_{i=1}^{n}(\omega-\omega_i)$$

反过来，这也表示：

$$|I-\rho W|=\prod_{i=1}^{n}(1-\rho\omega_i)$$

奥德表明W的ω_i可以在模型剩余部分被估计之前就找到。

回顾方差一致的情况下，经典线性回归模型中的对数似然函数为：

$$\ln L(\beta,\sigma^2)=-N/2\ln(2\pi)-N/2\ln(2\sigma^2)-(y-X\beta)'(y-X\beta)/2\sigma^2$$

相比之下，空间滞后模型的对数似然函数为：

$$\ln L(\beta,\sigma^2,\rho)=\ln|I-\rho W|-N/2\ln(2\pi)-N/2\ln(2\sigma^2)-(y-\rho W-X\beta)'(y-\rho W-X\beta)/2\sigma^2$$

并假定ω_i在估计之前就已知，这样我们就可以通过最大化该函数，而很容易地得到空间滞后y模型的最大似然估计值。我们还需要确认系数不会导致爆炸性的反馈过程，因为这将导致协方差矩阵非正定。

尽管当前存在其他算法，但我们在使用空间滞后y模型的最大似然法时仍然最常采用奥德的方法，该方法可以消除复杂计算中的一个主要部分。然而，在采取最大似然法时，基本假设却发生了变化。在OLS中，误差要求服从正态分布，但是不一定需要数据。而在空间模型的最大似然估计中，数据被假设服从正态分布。

第3节 空间性间隔 y 模型的最大似然估计：以民主研究为例

在本节中，我们将列出民主的空间滞后 y 模型的最大似然估计，同时将其与相同模型的 OLS 估计结果进行比较。

首先执行这个命令的 R 编码为：

```
sldv.fit <- lagsarlm(democracy ~ log(gdp.2002/population),
  data = sldv, nb2listw(nblist), method = "eigen", quiet = FALSE)
summary(sldv.fit)
moran.test(resid(sldv.fit),nb2listw(nblist))
```

上述命令的结果见表 2.3。正如我们所见，对人均 GDP 的系数估计(接近 1.0)高于 OLS 估计结果(0.76)，而空间滞后 y 的参数 $\hat{\rho}$ (0.56)却比空间滞后 y 模型的 OLS 估计更低。无论何种估计方法，我们的主要结论都是一致的，也就是：包括空间滞后 y 项将显著提高模型对国家间民主变化的解释能力。

如果我们相信最大似然估计 MLE 相比于 OLS 更适用于估计空间滞后 y 模型，那么我们可以推断：OLS 估计低估了人均 GDP 的系数，且高估了空间滞后项的系数。但这种

表 2.3 空间滞后 y 模型的最大似然估计

	$\hat{\beta}$	$SE(\hat{\beta})$	z Value
截距	−6.20	2.08	−2.98
人均 GDP 对数	0.99	0.28	3.59
$\hat{\rho}$	0.56	0.08	7.43
$N = 158$			
Log likelihood ($df = 4$) $= -491.10$			

推测并不可以检验，因为我们并不知道“真实”的参数是什么，以及我们的模型和真实的情况有多接近，甚至不确定是否存在“真实”参数值。

针对残差自相关的拉格朗日乘子检验，更常用于检验空间模型中的残差。在本例中该检验得到的值为 2.1，相应的概率为 0.147，也就是明确拒绝了剩余残差间的一阶自相关的情况。除此之外，衡量残差的空间聚集程度的莫兰 I 估计值得到了同样的结果，鞍点修正法结果也是如此。莫兰估计得到的标准分为−0.46，这表示基于同样的连接矩阵 W，我们拒绝了残差之间存在简单空间相关。如果我们对空间滞后 y 模型的 OLS 残差的空间模式进行检验，我们发现：大量证据表明残差之间仍然表现出很强的空间聚集效应，这时莫兰 I 为−0.17，相应的标准分为−3.21；鞍点估计结果也基本一致。负的莫兰 I 表明残差之间的排斥作用，这也支持我们的推测，即 OLS 过高估计了空间滞后 y 的作用，同时在对人均 GDP 效应的估计中过度矫正了空间依赖关系。我们强调应当谨慎使用残差的自相关检验，因为它们依赖于连接矩阵，而该矩阵本身在很多情况下都受到各种可能的识别方法的影响。我们在后文中还会讨论这一点。

第 4 节 | 空间滞后 y 模型的均衡效应

基于空间滞后 y 模型的最大似然估计，我们来考察人均 GDP 对民主影响的均衡效应。这需要我们考虑到自变量中一个国家 i 发生变化时，其他国家受到的影响。这将通过连接矩阵影响到其他国家的一系列变化，并最终通过空间滞后 y 项影响 y_i。

请记住，空间滞后回归模型可以被写成以下的矩阵形式：

$$y = X\beta + \rho W y + \epsilon$$

将所有和因变量 y 有关的项移到左边，我们得到：

$$(I - \rho W) y = X\beta + \epsilon$$

求解关于 y 的方程，再求期望值，我们可以发现，在这个均衡中，y 的期望值为：

$$E(y) = (I - \rho W)^{-1} X\beta$$

很明显，只有当 $\rho = 0$ 的时候，$E(y)$ 才可能缩减为 $X\beta$。为了确定 y_i 的期望值或者 x 的均衡效应，我们必须考虑到空间乘子 $(I - \rho W)^{-1}$。这个乘子告诉：我们 x_i 的变化多少会"扩散"(spill over) 到其他国家 j，并反过来通过 y 的空间滞后量影响到 y_i。这和列昂杰夫(Leontief, 1986)在投入产出

分析中应用的逆矩阵(inverse)方法相似，它用来估计一个部门的需求变化如何影响多个部门体系的总体产量。

因此，为了确定 x_i 中某个观测值的一个单位变化所带来的均衡效应，我们需要乘上向量 $\Delta x(i)$，而其他单位 j 的值都通过$(\mathrm{I}-\rho\mathrm{W})^{-1}\beta$控制为常数。由于每个国家和其他国家之间的连接程度不同，并且高阶连接程度也不同，因此对一定 x_i 变化产生的影响将根据特定国家的改变而改变。假设我们有两个不相连的区域，彼此之间不存在桥梁连接，那么区域 1 的改变将影响到区域 1 中其他国家的改变，但是这些改变对区域 2 中的国家将没有影响。

描述均衡效应变化的一个有用方法，是考虑所有国家发生一定变化时产生的影响，并且考察每个具体国家估计值的分布。在这个例子中，我们得到的均衡效应均值为 1.09，它大约比表 2.3 中 $\hat{\beta}$ 系数估计值 0.99 所表示的人均 GDP 对数的短期效应高 10%。在个体上具体国家的均衡效应从低的 1.03(蒙古)到高的 1.24(巴布亚新几内亚)，后者比人均 GDP 的短期效应高 25%。很明显，我们不应该在还没有考虑到空间乘子以及空间个体之间的变动时就对空间滞后 y 模型中的协变量的作用进行推断。图 2.1 展示了估计效应的柱状图。

考察全向量 $(\mathrm{I}-\rho\mathrm{W})^{-1}\beta\Delta x(i)$，将有助于我们理解一个国家的人均 GDP 如何影响其他国家的民主预期值。比如以俄罗斯为例。表 2.4 列出了基于俄罗斯得到的 10 个 $(\mathrm{I}-\rho\mathrm{W})^{-1}\beta\Delta x(i)$ 的最高值，表 2.3 列出了该空间滞后 y 模型的估计值，连接矩阵用 W 表示。如我们所见，俄罗斯的潜在均衡作用为 1.09，这与模型中潜在均衡影响的中位数接近。其

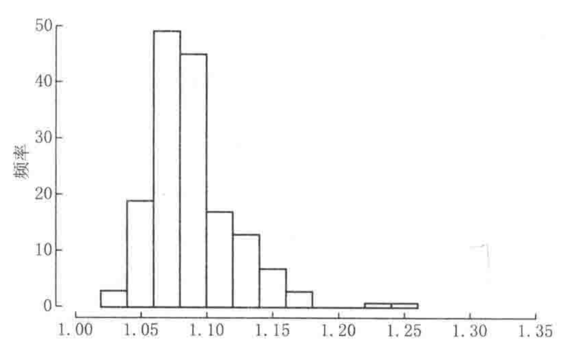


图 2.1 人均 GDP 自然对数均衡效应的柱状图

表 2.4 俄罗斯等 10 个最高人均 GDP 得分国家的均衡作用

国　家	作　用	国　家	作　用
俄罗斯	1.09	爱沙尼亚	0.21
朝　鲜	0.24	挪　威	0.20
日　本	0.24	立陶宛	0.20
蒙　古	0.24	拉脱维亚	0.12
芬　兰	0.22	亚美尼亚	0.18

他国家的值表明:俄罗斯的改变将会对其亚欧邻国产生影响。为了弄清这些估计值在实际中表示的含义,我们可以重新回忆一下人均 GDP 对数的估计效应的系数。俄罗斯当前 GDP 每变化 10%(即 2279 美元),将仅仅使其民主预测值的 POLITY 得分增加 0.1 分。在和俄罗斯发生相同的人均 GDP 变化情况下,即便是在该变化所产生的均衡效应最大的国家,其民主预测值也仅仅在此估计值基础上增加 0.02。这进一步强化了我们的结论:即便某个国家的人均 GDP 发生很大差异,但根据我们的模型,这并不会使世界范围内的其他国家的民主预期水平产生很大的改变;同时,当空间滞后 y 模型考虑了相连国家之间民主水平的相互影响之后,人均 GDP 对数的影响将远远小于把各个观测值看做相互独立时

的 OLS 结果。

根据上文的空间滞后 y 变量估计值，我们利用下面的编码建构一个简化实验：

```
# Code to calculate equilibrium effect of changes in GDP per capita
# Create vector to store the estimate for each state
ee.est <- rep(NA, dim(sldv)[1])
# Assign the country name labels
names(ee.est) <- sldv$tla
# Create a null vector to use in loop
svec <- rep(0, dim(sldv)[1])
# Create a N x N identity matrix
eye <- matrix(0, nrow = dim(sldv)[1], ncol = dim(sldv)[1])
diag(eye) <- 1
#Loop over 1:n states and store effect of change in
# each state i in ee.est[i]
for(i in 1:length(ee.est)){
        cvec <- svec
        cvec[i] <- 1
        res <- solve(eye - 0.56315 * wmat) %*% cvec * 0.99877
        ee.est[i] <- res[i]
}
#Russia example of impact on other states (observation 120)
cvec <- rep(0, dim(sldv)[1])
cvec[120] <- 1
# Store estimates for impact of change in Russia in rus.est
eye <- matrix(0, nrow = dim(sldv)[1], ncol = dim(sldv)[1])
```

```
diag(eye) <- 1
rus.est <- solve(eye - 0.56315 * wmat) %*% cvec * 0.99877
# Find ten highest values of rus.est vector
rus.est <- round(rus.est, 3)
rus.est <- data.frame(sldv$tla, rus.est)
rus.est[rev(order(rus.est$rus.est)),][1:10,]
```

前面OLS模型的结果表明:人均GDP对民主预期水平的影响是相对有限的。表2.3中空间滞后 y 模型的最大似然估计结果同样表明人均GDP的即刻影响相对较小。当我们考察人均GDP变化的长期均衡效应时,尽管该效应略微大但仍然有限。表2.3中的系数意味着一个国家的预期民主水平和其邻国之间存在什么样的关系呢?图2.2显示出模型中地理上预期的共变关系。在该图中,我们画出了因变量(民主,y)的期望值和其邻国之间民主水平(空间滞后变量 y^s 和自变量人均GDP的自然对数)的函数关系。该等高线

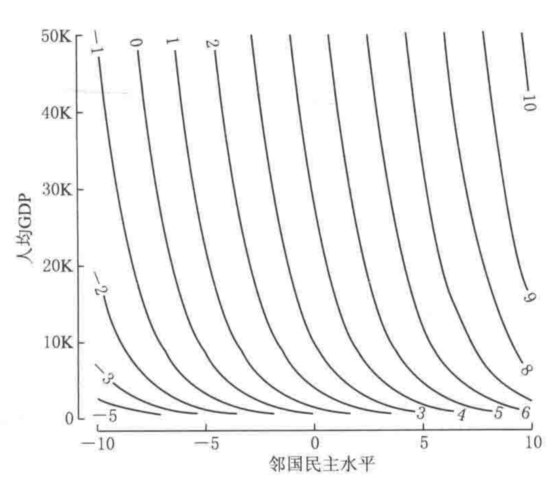


图2.2 根据邻国民主水平和人均GDP对数预计得到的民主水平变化图

图明确表明：人均 GDP 的作用很小，但是空间组成对国家的预期民主水平具有很大的作用。

模型表明：控制了 GDP 水平后，一国的民主期望水平出现了很大的差异。如果一个国家的人均 GDP 取中位数，其邻国全部为专制国家（比如 $y_i^s = -10$），其期望民主得分约为 -4；然而如果它的所有邻国为民主国家（比如 $y_i^s = 10$），那么它的期望民主水平将接近 7。在这种情况下，尽管人均 GDP 在解释民主差异上的作用非常有限，但是国家的民主水平与其邻国之间却具有非常紧密的关系。

针对该结果的另外一种思路是，如果民主的改变来自于没有包含在模型系统部分中的其他特征（比如某个国家 i 发生的震动）以及随之产生的对其他国家 j 的预测民主水平（在模型中表示为 $\hat{y}_j$）的短期影响，那么会出现什么结果呢？基于估计出的空间模型，这将对其他国家产生什么影响？很明显，这不仅依赖于 $\hat{\rho}$，而且还和建构空间滞后性的连接矩阵 W 的结构有关。

基于上面的空间滞后 y 估计，我们利用下面的编码来考察中国一个单位的 y 变化对系统中其他国家所产生的影响：

```
# Impact of change in $y$ to 10 in China
# China is observation 32
cvec <- rep(0, dim(sldv)[1])
cvec[32] <- 10
# Store estimates of change in China in chn.est
chn.est <- c(cbind (0, 0, wmat %*% cvec) %*%
```

```
    c(summary(sldv.fit)$Coef[, 1],summary(sldv.fit)$rho))
chn.est <- round(chn.est, 3)
# Find all states where non-zero impact
chn.est <- data.frame(sldv$tla, chn.est)
chn.est <- chn.est[rev(order(chn.est$chn.est)),]
chn.est[chn.est$chn.est>0,]
```

表 2.5 如果令中国的 POLITY 得分为 10,预测民主 $\hat{y}$ 的作用

国家(或地区)	作用	国家(或地区)	作用	国家(或地区)	作用
中国台湾	1.88	老 挝	1.13	塔吉克斯坦	0.80
朝 鲜	1.88	吉尔吉斯斯坦	1.13	印 度	0.80
蒙 古	1.88	孟加拉国	1.13	越 南	0.80
尼泊尔	1.41	乌兹别克斯坦	0.94	阿富汗	0.80
不 丹	1.41	泰 国	0.94	哈萨克斯坦	0.70
巴基斯坦	1.13	缅 甸	0.94	俄罗斯	0.28

第5节 | 意大利投票率的空间依赖关系

锡恩(Shin, 2001)以及锡恩和阿格纽(Shin & Agnew, 2002、2007a、2007b)研究了意大利在过去的几十年中其政治活动的地理分布,并且发现在投票率和选举结果之间存在重要的空间动态关系。我们利用他们的数据来解释一种简单的观点,即投票率的空间变化可以通过意大利财富与收入的地理分布来解释。数据来自意大利2001年国家选举和各省在1997年的人均GDP。这些数据包括选举中所有的477个选区(collegi),或者叫单一席位选区(后面将用SMDs来表示)。

主要变量的地图

与空间回归分析中的第一步一样,这里我们先画出了选举投票率和人均GDP的地理分布关系(见图2.3和图2.4)。

投票率最高的是北部,尤其是北部最远的米兰(Milan)附近以及艾米利亚—罗马涅(Emilia-Romagna)和托斯卡纳(Tuscany)。罗马(Rome)和威尼斯(Venice)的投票率也很高。例如,在摩德纳(Modena),投票率为90%。相反,在西西里(Sicily),投票率就停留在百分之十几;即便是在那不勒

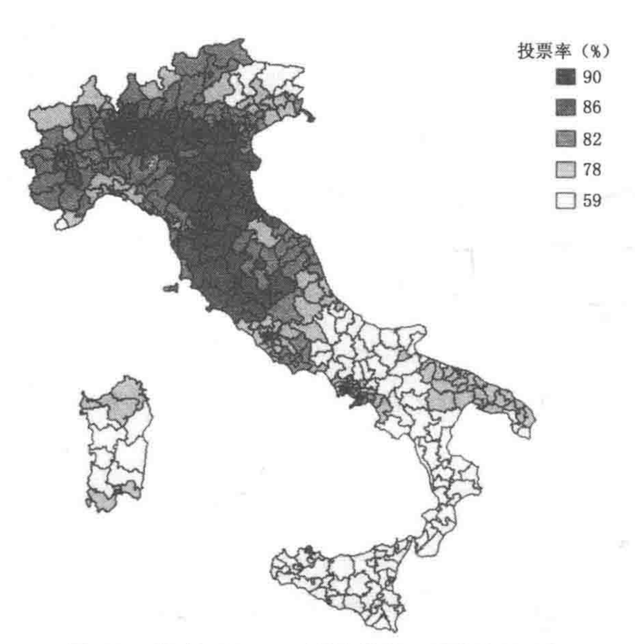


图 2.3　按选区(Collegio)划分的意大利选举投票率

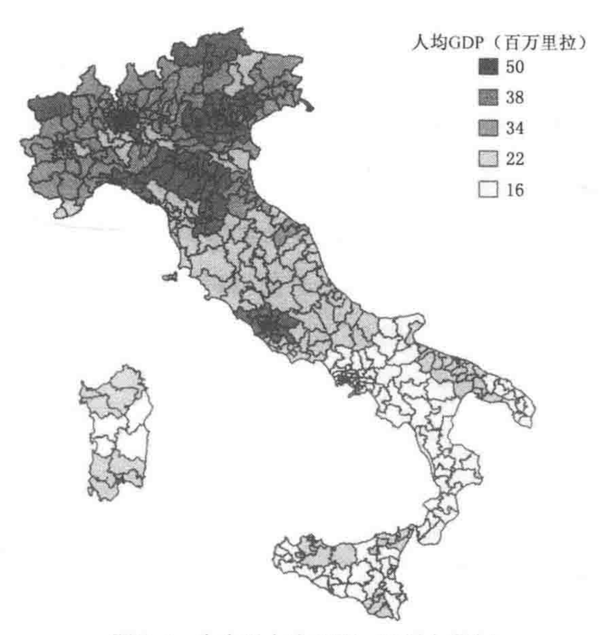


图 2.4　意大利人均 GDP，1997 年数据

斯(Naples)市郊,选举投票率也仅仅约为60%。从人均GDP来看,意大利最富裕的地区是伦巴底(Lombardy)。北部最富裕选区的收入大约是南部最穷选区收入的1.5倍。这组探索性地图明显表现出了投票率和人均GDP的聚集效应。

计算莫兰I统计量

在本节中,我们将利用莫兰I统计量这种更正规的方式来测算投票率和人均GDP的空间聚集效应。在测量意大利空间连接的第一步,我们先选取50千米作为最近的邻近距离。通过计算各个区的图形中心,由此判断该区是否和其他区域的中心在50千米范围之内。总的看来,有两个米兰区域和54个其他区域相连:十区和六区。另外存在八个区仅仅和另外一个单一区相连,但这主要是边缘效应:比如特伦提诺阿托(Trentino-Alto Adige),就位于阿尔卑斯山勃伦那(Brenner)山口和奥地利边境上。然而总体来说,各选区平均和其他17个相隔50千米以内的区相连。

对临近区域一览表的概括分析可以通过R很容易地得到。作为例子,我们用如下编码生成这些连接:

```
tr <- readShapePoly("turnout",
  IDvar = "FID_1", proj4string = CRS(" + proj = robin  + lon 0 = 0"))
dnn50km <- dnearneigh(coordinates(tr), 0, 50000)
summarize(dnn50km)
```

这里有两种莫兰计算方法,一种是基于随机假设,另外

一种是基于正态假设。不管我们是基于何种假设，莫兰I统计量都表明数据中存在很强的空间关系。在这两种检验中，我们发现人均GDP的莫兰I为0.86；类似的，投票率的相关上也具有很高的取值，即0.79(两者都是)。所有这些值都是异常的，也就是表明人均GDP和投票率之间都存在很强的空间关联模式。

回归分析

投票率和人均GDP之间明显可能存在相关关系，但是投票率的空间聚集关系可以完全被GDP的地理差异解释吗？下面考察了投票率和人均GDP的简单函数模型。首先我们考查标准的最小二乘估计，结果在表2.6中列出。标准结果表明：在意大利，收入是投票行为的一个很强预测指标，对数人均GDP一个单位的变化(每百万里拉)将导致14%的投票率变化。然而，我们得到回归残差空间模式的莫兰值为0.47，这表明还存在没有被协变量解释掉的空间模式。

表2.6 意大利投票率对人均GDP对数的OLS回归(1997)

	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距	35.30	2.21	15.96
人均GDP对数	13.46	0.65	20.84
$N=477$			
Log likelihood ($df=3$) $=-1387.57$			
$F=434.4(df_{1_}=1, df_{2_}2=475)$			

然后我们用下面的代码，考察空间滞后y回归模型：

```
shin <- read.csv("italyturnout.csv",sep = ",",header = T)
sldv.fit <- lagsarlm(turnout ~ log(gdpcap), data = shin,
      nb2listw(dnn50km), method = "eigen", quiet = FALSE)

summary(sldv.fit)
```

表 2.7 列出了计算结果。人均 GDP 对投票率的作用没有上面的 OLS 结果那么“强”,但是更可信。它表明收入的作用减小,但仍然是一个很强的作用。然而,空间滞后变量的作用相当显著。

表 2.7 意大利投票率对人均 GDP 对数的空间滞后回归(1997)

	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距	4.70	1.66	2.80
人均 GDP 对数	1.77	0.48	3.66
$\hat{\rho}$	0.87	0.02	36.7
$N = 477$			
Log likelihood ($df = 3$) $= -1193$			

均衡分析

根据上面的方法。可以很容易地算出 477 个选区中每一个的均衡值,也就是模型的预期值。这里我们并没有将它们列出来,而是用一个简单的实验方式:假设意大利最穷区域雷焦卡拉布里亚—斯巴里(Reggio Calabria-Sbarre)的人均 GDP 翻倍,这样我们计算出这一“情景”下的预期值和模型中观测数据的预期值的差异。这个差异对大部分选区来说都

是不存在的，但是对其附近的15个选区来说，这种单一选区人均GDP的变化将导致期望投票率发生1%或者更大的改变。与预期一样，最大的改变是在相邻的选区之间。图2.5表示了意大利投票率发生变化后的分布。



图2.5 由于南部意大利单一穷困区（雷焦卡拉布里亚—斯巴里）人均GDP翻倍而带来的预期投票率增长

下面的编码表示如何基于事先得到的空间滞后 y 值，构造针对雷焦卡拉布里亚—斯巴里（432个观测值）的实验（简略形式）：

```
# Extract estimated rho
rho <- coef(sldv.fit)[3]

# Extract estimated beta
```

```
beta <- coef(sldv.fit)[1:2]

# Create a X matrix
X <- cbind(1,log(shin$gdpcap))

# Create an alternative X matrix, changing value for
# Reggio Calabria - Sbarre (obs 432)
Xs <- X
Xs[432]<- log(35)

# Create an identity matrix
I <- diag(length(shin$gdpcap))

# wmat is the weights matrix
wmat <- nb2mat(dnn50km,style="W")

# Find equilibrium effect by looking at
# the difference in expected value for the
# two x matrices
Ey <- solve(I - rho*wmat)%*%(X%*%beta)
EyS <- solve(I - rho*wmat)%*%(Xs%*%beta)
dif <- Eys-Ey
```

将不同权重矩阵引入空间滞后因变量模型中

我们以 2004 年美国总统选举为例，说明空间权重矩阵的作用[10]。从该数据中，可以很容易提取出 XML 表，并且转变成 csv 文档。为了简化研究问题，我们将不考虑阿拉斯加和夏威夷，因为它们和其他所有的州都隔得非常远，所以

并不会影响我们对区域数据的分析。我们感兴趣的主要变量，是小布什和克里的总选票在 48 个相连州与哥伦比亚地区中的份额。考虑到这是作为练习的一个例子，我们忽略每个州的提名投票数。然后构造一个小布什对克里的投票比，将此作为因变量。

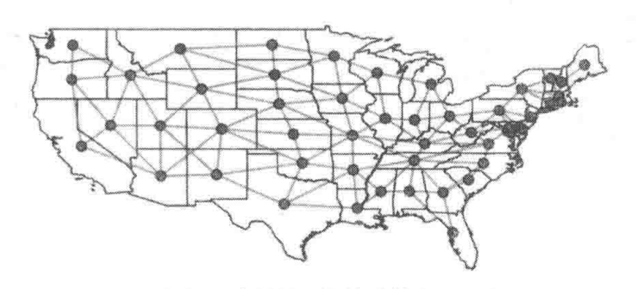


图 2.6 美国 48 个州按照接壤计算出的一阶连接图

为了找出空间类型数据中的自相关程度，我们构造出这 49 个政治和地理单位之间空间连接性的几种测量。第一种测量是相连州之间的简单测量。在这种情况下，华盛顿州与爱达荷州和俄勒冈州相邻，因为它们接壤。科罗拉多州与新墨西哥州、亚利桑那州、犹他州、怀俄明州、内布拉斯加州、堪萨斯州以及俄克拉荷马州接壤。在分布的另一端，缅因州只与一个州接壤。图 2.6 描述了这些情况。

绘制这个地图的编码样本如下：

```
library(maptools); library(network)
library(spdep); library(sp); library(rgdal)
setwd("...")

# read in 2004 presidential votes
```

```
presvote <- read.table("2004presvote.csv", sep = ",", head = T)

# read in shape files for 48 US States plus District of Columbia
# will create a MAP OBJECT
# use equal area projection(Robinson)
usa.shp <- read.shape("48_states.shp")

usaall <- merge(usa.shp$att.data, presvote,
        by.x = "STATE_NAME", by.y = "State",
        sort = F)

# Create a distance matrix from original polygon shape file
tr <- readShapePoly("48_states, shp",
  IDvar = "ObjectID", proj4string = CRS(" + proj + robin  + lon 0 = 0"))
centroids <- coordinates(tr)

# Create polygons in a spatial object
us48polys <- Map2poly(usa.shp,
        region.id = as.character(usa.shp$att.data$STATE_NAME))

# Create neighbors, list, and matrix objects from polygon
  centroids
us48.nb <- poly2nb(us48polys,
        row.names = as.character(usa.shp$att.data$STATE_NAME))
us48.listw <- nb2listw(us48.nb, style = "B")
us48.mat <- (nb2mat(us48.nb, style = "B"))

# plot the network among the centroids
colnames(us48.mat) <- rownames (us48.mat) <- usa.shp$att.dat
$STATE_ABBR
usa <- network(us48.mat, directed = F)
```

```
set.seed(123)
# plot network first; then add state boundaries
plot.network(usa1, displayisolates = T, displaylabels = F,
      boxed, labels = F, coord = centroids, label.col = "gray20",
      usearrows = F, edge.col = rep("gray60", 190),
      vertex.col = "gray30", edge.lty = 1)
plot(us48polys, bty = "n", border = "slategray3", forcefill = TRUE,
    xaxt = "n", yaxt = "n", lwd = .000000000125, las = 1,
    ylab = "", xlab = "", add = T)
```

下面我们用地图画出 2004 年总统选举中，小布什和克里在每个州的得票比。如图 2.7 所示，2004 年总统大选中各州之间的投票行为表现出很强的地理关系。

```
library(RColorBrewer)
# now plot the Bush: Kerry vote ratio
bk <- usaall$Bush/usaall$Kerry
# set up five categories and assign colors
breaks <- round(quantile(bk), seq(0, 1, 1/5), na.rm = TRUE), 1)
cols <- brewer.pal(length(breaks), "Greys")

# use findInterval to color states by bk variable
plot(us48polys, bty = "n", border = "slategray3", forcefill = TRUE,
    xaxt = "n", yaxt = "n", lwd = .000000000125, las = 1, ylab = "",
    xlab = "")
plot(us48polys, bty = "n", col = cols[findInterval(bk, breaks,
all.inside = T)], forcefill = T, add = T)
```

```
legend(x = c( - 125, - 115), y = c(27,32), legend = leglabs
(breaks),
        fill = cols, bty = "n")
```

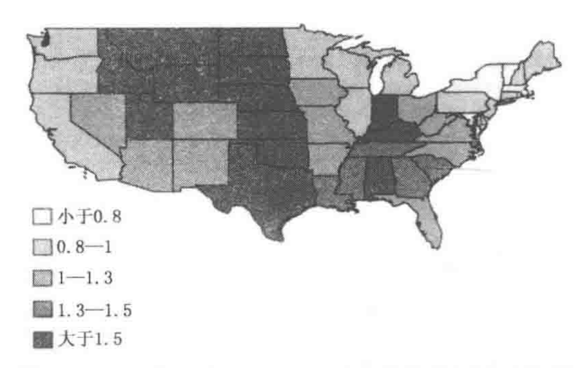


图 2.7　2004 年总统选举各州支持小布什对克里的投票比

莫兰I同样在数字上表明了这种空间关系(见表 2.8)。

表 2.8　2004 年总统选举小布什—克里总投票数的自相关

数　　量	标准分	加权模式
莫兰(Moran's) I		
0.39	4.7	边境接壤的州
0.49	5.7	最近的 4 个邻州
0.30	7.0	最近的 12 个邻州
吉尔里(Geary's)C		
0.65	−2.7	边境接壤的州
0.65	−3.6	最近的 4 个邻州
0.69	−5.1	最近的 12 个邻州

各州的平均生产总值(类似于 GDP)来自于经济分析局的测量,该局是美国商务部的一部分。最新的可用数据来自于 http://www.bea.gov/bea/newsrelarchive/2006/gsp1006.xls,这也包括 1997 年到 2004 年各州生产总值的增长率。这

些数据描绘了在 2004 年选举之前的七年中，各州经济的变化。我们用该数据作为协变量来解释 2004 年总统选举的投票情况。

我们建立了两个基本的空间连接矩阵，一个是根据是否接壤，另外一个是选择最近的四个邻州。对于每一个接壤编码，我们都估计了一个空间滞后变量的回归模型。结果见表 2.9。

表 2.9 2004 年美国总统竞选中各州小布什对克里的投票率对 GDP 增长率(1997—2004 年)的空间滞后回归

No.	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
是否与皇后地区(Queen)接壤			
截距	0.86	0.21	4.00
GDP 增长率	−0.05	0.06	0.85
$\hat{\rho}$	0.09	0.02	20.4
$N = 49$			
Log likelihood ($df = 3$) $= -25.63$			
4 个最接近的邻近州			
截距	0.63	0.23	2.72
GDP 增长率	−0.06	0.05	1.04
$\hat{\rho}$	0.60	0.12	18.4
$N = 49$			
Log likelihood ($df = 3$) $= -25.19$			

经验研究结果表明，GDP 增长率更高的州，在投票上更少支持小布什而更多支持克里。在接壤编码上，我们发现了很弱但正向的空间关系(0.09)，但在利用四个邻州作为空间加权编码的时候，小布什—克里投票比却表现出更强的正向相关关系(0.60)。这两种不同的估计方式不仅在标准回归结果表上出现不同的结果，而且更重要的是它们将导致均衡值出现很大差异。如图 2.8 所示，对于这两种不同的加权模式，尽管其均衡效应正向相关，但它们的分布却完全不同。

边境接壤方法得到的均衡效应的均值(−0.15)比用四个邻近州作为估计得到的均值(−0.35)更小。该例子说明的首要问题是:加权矩阵在空间分析中起到了重要的作用,即便加权矩阵中很小的波动,也会对经验结果产生显著的影响。

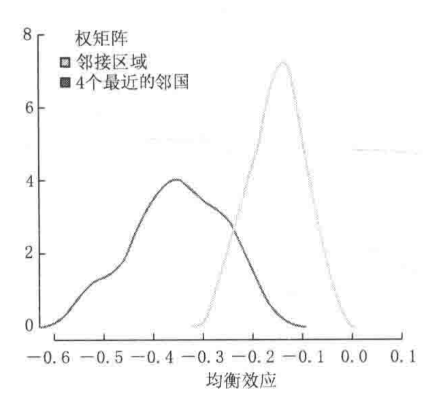


图 2.8 不同加权模式下空间滞后因变量的均衡效应的后验分布

空间滞后因变量与 OLS 中虚拟变量的比较

社会科学家们经常意识到:世界各个地区之间存在巨大的异质性,因此回归模型中每个国家各自的协变量并不能很好地解释空间差异。解决空间异质性的通常做法,是将各个不同地区作为虚拟变量加入模型。这种虚拟变量可以完全拟合不同地理区域的截距,从而考虑各个分散区域中因变量 y 的固定均值差异。这是在当前应用研究中最常用的处理区域异质性的方法,社会科学中有很多模型都是这样将"区域"分类作为虚拟变量加入模型的。同时,这样的模型也变得越来越普遍,因为分析者们越来越多地注意到合并数据的 OLS (pooled OLS)估计可能没有考虑到各区域的重要差异。

比如，李(Lee, 2005)在一项民主和公共部门规模对收入不平等的影响的研究中，将非洲、亚洲和拉丁美洲作为区域虚拟变量拟合模型，并提出后两个区域相比于参照类(经济合作和发展组织)明显不同，并且这种不同不能通过模型右边国家方面的具体变量来解释。在民主研究中，布赫特(Burkhart)和刘易斯—贝克(Lewis-Beck, 1994)将世界体系中不同的民主水平作为虚拟变量，以此处理模型中的异质性问题，从而将国家区分为世界经济的中心(core)、边陲(periphery)和半边陲(semi-periphery)。

在模型中加入区域虚拟变量的方法在社会科学中非常受欢迎，这也成为空间滞后 y 模型的一种替代方法。下面我们将区域虚拟变量加入模型中，作为原始 OLS 模型的替代方法，然后讨论这种模型和空间滞后 y 模型的关系。表 2.10 列出了将拉丁美洲和加勒比海、欧洲、撒哈拉以南非洲地区、中东和北非、亚洲以及大洋洲作为虚拟变量的模型。该模型忽略的区域，也就是参照类，为北美洲(即美国和加拿大)。不同区域的系数估计表明：在控制了人均 GDP 以后，某一地区的一个国家预测的民主水平相比于北美的差异。拉丁美洲和加勒比海地区、欧洲、大洋洲看上去和北美的平均民主水平之间不存在显著差异，但是撒哈拉以南非洲地区和亚洲，尤其是中东，相比于北美，其平均民主水平要低很多。我们同时注意到：这里的人均 GDP 对数的估计系数要比将各个国家作为独立个体得到的 OLS 模型(也就是 1.68)低很多。事实上，该模型的人均 GDP 系数和上面空间滞后 y 模型中得到的均衡效应(也就是 1.09)非常接近。这也表明合并数据的 OLS 忽略了空间异质性，同时通过虚拟变量控制区

域差异，起到了解决区域异质性的作用。此外这也说明之前人均 GDP 的作用被高估了。

表 2.10 包含空间虚拟变量的模型估计

No.	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距	−1.89	5.06	−0.37
人均 GDP 对数	1.15	0.34	3.39
拉丁美洲和加勒比海地区	0.09	3.84	0.02
欧洲	−0.41	3.74	−0.11
撒哈拉以南的非洲地区	−4.71	3.97	−1.19
中东和北非	−11.77	3.85	−3.05
亚洲	−5.97	3.92	−1.52
大洋洲	0.90	4.72	0.19
$N=158$			
Log likelihood ($df=8$) $=-477.52$			
$F=18.65$ ($df_1=7$, $df_2=150$)			

空间虚拟变量的方法适合替代空间滞后 y 模型吗？回答这个问题的一个可能的办法是考虑这两个模型的简约性。尽管包含虚拟变量的 OLS 具有更高的对数似然比，但这增加了六个新的参数，或者说比空间滞后 y 模型多出了五个参数。同时，区域虚拟变量模型本身并没有解释为什么会存在区域差异，它仅仅是基于观察到的区域差异拟合不同的截距。如果一个区域因为其人均 GDP 变化而导致其民主水平发生了改变，这也不会改变其他国家的预测值，因为区域差异被看做固定的，并且国家之间也不会相互影响。相比之下，空间滞后 y 模型仅仅增加了一个参数，就可以从实质上解释与某一国家相连的国家的民主水平 y 对这个国家民主水平的作用。如果不考虑简约性，而仅仅是为了提高拟合度，当然可以用区域虚拟变量拟合空间滞后 y 模型。但在这种情况下，模型仍然表现出残差的空间聚集效应，也就是得

到正向且统计显著的估计值 $\hat{\rho} = 0.25$。如果用空间虚拟变量拟合空间滞后 y 模型，那么分析者需要考虑的，是固定区域差异的相关假设问题，而不是空间滞后 y 表达式中所产生的内生性问题。在很多情况下，如果连接矩阵 W 和区域划分非常相似的话，我们将很难分别估计区域虚拟变量的参数和空间滞后 y 的参数，这与回归的共线性问题类似。进一步而言，区域虚拟变量的构造基于指定国家为分散无关联的或者特定名称的区域，而空间滞后 y 项的构造基于连接矩阵 W，每个国家都有特定的连接。与分散无关联的识别方法相比，基于国家之间连接的方法的优势在于，它不会将地理上隔得很远的国家比如希腊和爱尔兰划为同一个群，也不会将跨越了几个共同的地理区域的国家，比如土耳其和俄罗斯，作为同一个区域处理。

即使我们相信互斥的区域也可以用来构造观测值之间的连接，但区域虚拟变量方法并不总是空间滞后 y 模型的合适替代，而且它需要更多过度限制的假设。为了说明这个问题，假设一个回归中有 k 个不同虚拟变量 D_k 且满足 $y = b_1D_1 + \ldots + b_kD_k + e$。与连接列表或者连接矩阵不同的是，在连接表或者矩阵中，i 不作为其自身的邻国包含在内，而在这里，每个区域同时包括 i 和它所有的邻国。但是如果每个区域内的样本数很大，我们可能得到 $Wy \approx b_1D_1 + \ldots + b_kD_k$。这表示虚拟变量回归可以重新写作 $y = b_1D_1 + \ldots + b_kD_k \approx Wy + e$。这样，虚拟变量回归模型就变成空间滞后 y 模型的特例，这仅仅是假设 $\rho = 1$ 而不是估计实际参数就是 1 (Lin et al., 2006)。换句话讲，空间虚拟变量假设每个区域内的所有的观测值都是同质并且相互连接的，但是空间滞后

y 模型却允许我们估计不同的相似程度。此外，空间滞后 y 模型可以方便地处理各种连接形式，而虚拟变量方法则假设群体之间互不相连。也就是说，群体内的每个分析单位相互关联，而群体之间没有关联；另外，分析单位也不能同时属于不同的群体。

第3章

空间误差模型

在第2章中，我们考察了空间滞后因变量模型，在该模型中因变量的“邻近”值对它本身具有直接影响。尽管这可能是处理空间依赖问题最常见并且是最有用的方法，但是，在连续因变量线性模型中，这并不是表示空间依赖关系的唯一方法。在本章中，我们将考察另一个替代概念，也就是说空间依赖关系来自于误差，而不是来自模型的系统部分。这种模型通常被称之为空间误差模型。我们同时介绍空间回归模型的一种重要扩展方法，也就是将空间误差模型扩展到度量学(metric)的距离概念，而不是仅仅停留在地理距离上。

第1节 | 空间误差模型

尽管空间滞后变量模型将空间依赖看做本质现象，也就是说 y_i 受到其他国家 $y_j(i \neq j)$ 的值的影响，但是，空间误差模型却将空间相关关系看做一种干扰。这类似于统计学方法中经常将时间序列相关视为需要消除的问题，而且仅仅是一种估计问题。这种方法一般关注模型系统部分中被估自变量的参数，而忽略了数据产生过程中，数据相关性本身所反映出的意义。空间误差模型假设模型的误差是空间上相关的，而不是认为 y_i 对 y_j 产生直接影响。这种空间相关性的建立可以通过很多种方法来识别。这里我们仅仅关注一种基于空间体系编码的简单方法，也就是空间权重的方法；其他重要的方法包括考察地理统计协方差结构，但是本书不介绍这些方法。根据之前的定义，如果令 w_i 代表向量 W，也就是测量 i 和其他单位 $j \neq i$ 的接近程度，我们可以将空间误差模型写成如下的形式：

$$y_i = x_i\beta + \lambda w_{i\cdot}\xi_i + \epsilon_i$$

这里我们将总的误差分解为两部分——即：ϵ 为空间不相关项，它满足正常回归假设中误差项空间上不相关的条件，ξ 为包含空间因素的误差项。参数 λ 表示在连接向量 w_i 中，相

邻观测值的空间误差项 ξ 的相关程度。另外，我们还可以将空间误差模型按照第 2 章中定义的项，写成矩阵的形式：

$$y = X\beta + \lambda W\xi + \epsilon$$

$$\epsilon \sim N(0,\ \sigma^2 I)$$

如果相连观测值 i 和 j 的误差之间没有空间相关存在，那么空间误差参数 λ 将为 0，同时模型将简化为标准线性回归模型，也就是个体观测值之间相互独立，这样我们就可以按照传统方法估计 OLS 模型。然而，如果空间误差参数 $\lambda \neq 0$，那么我们可以判断相连观测值的误差存在空间依赖。这种结果可能仅仅是出于巧合，或者反映了模型系统成分的识别错误，尤其当被省略变量存在空间聚集的时候。社会科学家们通常希望发现正向空间相关关系。这意味着相似值之间存在聚集；也就是观测值 i 的误差的大小会随着其邻近观测值 j 的误差变化而系统性变化，这样 i 的更小/更大的误差就会向 j 的更小/更大[①]的误差靠拢。这种残差聚集效应违背了误差项相互独立的假设。

误差项的空间相关将导致什么结果呢？如果我们按照误差项相互独立的假设来估计 OLS，又会导致什么结果呢？如果 $\lambda \neq 0$，即便忽略了空间相关关系，OLS 系数估计结果仍然是无偏的。然而，系数估计的标准误将会是错误的。在 OLS 对方差的估计中，假设观测值相互独立。如果这不正确的话，那么，OLS 对方差 $\hat{\sigma}$ 的估计将低估了实际的方差，这与时间序列相关情况下的误差估计类似。这种情况之所以发

① 原书这里为更小，而译者认为 i 的误差与 j 的误差大小应该是对应的。——译者注

生，是因为对方差的估计忽视了相邻观测值误差项的相关。此外，估计系数也并不必然是我们想要得到的有效估计值，也就是不“接近”真实值。之后，我们将返回来讨论空间误差模型的估计，下面我们先介绍如何解释该模型以及它和空间滞后 y 模型的关系。

空间误差模型和空间滞后 y 模型表面上看起来很相似，因为它们都表明了观测值之间的空间依赖性。然而，实际上这两种模型识别方法的实质意义却是不同的。空间滞后 y 模型是一种联立模型（simultaneous model），观测值之间相互进行反馈：y_i 的值影响到 y_j 的值，反过来影响到 y_k 的值，然后再影响到 y_i。正如在第 2 章中所见，一个观测值 i 的自变量的不同取值，将通过相连的观测值而传播，而净影响来自于这些不同取值通过空间滞后 y 项，对其他相连观测值产生的影响。与之相反，在空间误差模型中，模型识别中的依赖关系仅仅来自于误差项。空间滞后 y 项的缺失表明 i 的自变量的差异并不会影响到与 i 相连的其他观测值的结果。因此，在空间误差模型的识别中，观测值的相互联系仅仅是因为未测量到的因素，也就是因为某些未知因素在距离上相关。

第2节 | 空间误差模型的最大似然估计

在空间滞后 y 模型的例子中，方程右边代表空间滞后作用的系数 ρ 是明确表明我们研究兴趣所在的参数。在空间误差模型中，λ 系数表明残差的相关，而不是明确的研究兴趣的协变量。如果我们仅仅是对估计 x 的 $\hat{\beta}$ 感兴趣，而完全忽略 λ，OLS 估计值将会是空间误差模型的无偏和一致性估计，这与空间滞后 y 模型不同。然而，报告的标准误却是不正确的，同时估计系数也不一定是有效的。这种问题可以通过利用广义最小二乘估计来解决，这类似于在存在时间相关的情况下用广义最小二乘估计值，这样我们先估计序列相关，然后试图转换数据以清除序列相关，从而满足普通回归的假设。通常这可以通过对空间连接矩阵特征值的最大似然估计来解决。

空间滞后误差模型的对数似然方程为：

$$\ln L(\beta, \sigma, \lambda) = \ln|\mathrm{I}-\lambda\mathrm{W}| - N/2\ln(2\pi) - N/2\ln(\sigma^2) - (y-\lambda\mathrm{W}y-\mathrm{X}\beta+\lambda\mathrm{WX}\beta)'(y-\lambda\mathrm{W}y-\mathrm{X}\beta+\lambda\mathrm{WX}\beta)/2\sigma^2$$

如同空间滞后 y 模型的对数似然值一样，我们遇到如何计算行列式 $|\mathrm{I}-\lambda\mathrm{W}|$ 的对数的难题，这是一个难以估计的 n

阶多项式。然而，我们可以在此依据奥德（Ord, 1975）的结果将这个行列式写成连接矩阵 W 的特征值 ω_i 的乘积：$|I-\lambda W| = \prod_i (1-\lambda\omega_i)$。由于特征值 ω_i 可以在最优化之前决定，该步骤可以同其他参数的似然估计分开（Anselin, 1988; Bivand, 2002）。这种估计可以通过常用的软件选项来执行，包括 R 中的 spdep。

第 3 节 | 以民主和发展研究为例

为了说明一个空间误差模型应用的实际例子,首先我们重新讨论第 2 章中关于民主和财富的例子。我们利用和第 2 章相同的数据,在变量的所有构造上面都参照前面的细节。表 3.1 展示了民主和收入例子中的三组估计。第三列的估计结果表明考虑到空间相关误差后的模型估计结果,而第一列和第二列重复了第 1 章和第 2 章中 OLS 与空间滞后 y 模型的估计结果。

估计空间误差的 R 编码非常简单明了:

```
# data and variables as employed in chapter 2.
sem.fit <- errorsarlm(democracy ~ log(gdp.2002/population),
      data = sldv, nb2listw(nblist), method = "eigen", quiet = FALSE)
summary(sem.fit)
logLik(sem.fit)
```

如表 3.1 所示,空间误差模型的人均 GDP 对数的估计系数远远大于在空间滞后 y 模型中的结果,尽管还是没有不包含空间关系的 OLS 模型的结果大。直觉告诉我们 OLS 模型可能因为没有考虑到民主和人均 GDP 在相邻国家间的空

表 3.1 民主和人均 GDP 对数

变 量	OLS			SLDV			SEM		
	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截 距	−9.69	2.43	−3.99	−6.20	2.08	−2.98	−7.49	3.07	−2.44
人均 GDP 对数	1.68	0.31	5.36	1.00	0.28	3.59	1.39	0.38	3.66
$\hat{\rho}$				0.56	0.08	7.43			
$\hat{\lambda}$							0.58	0.08	7.60
N		158			158			158	
自由度(df)		1			2			2	
对数似然值 (Log likelihood)		−513.62			−491.10			−491.53	

注:SEM 表示空间误差模型;SLDV 表示空间滞后因变量。

间聚集作用，而高估了人均 GDP 的直接作用。这样估计结果也更不精确。同时，我们也可以将空间滞后项看作当观测值相互独立的情况下，OLS 模型中被忽略的变量。相比之下，空间误差模型纠正了人均 GDP 和民主的正向空间相关，而且这种纠正减小了 GDP 影响的估计系数。然而，空间误差估计假设模型中观测值的空间依赖仅仅来自于误差，或者说是模型系统部分没有考虑到的因素。

相反，在空间滞后 y 模型中，一国人均 GDP 增长带来的净效应，一部分会通过反馈效应实现，因为 i 的即时效应对其邻国 j 产生影响，然后通过空间滞后项又影响 i，这带来的民主分数的变化又会影响其他国家，并通过系统中的反馈直到产生某种均衡。因此，在空间滞后 y 模型中，人均 GDP 的估计系数看上去比空间相关误差模型更小，因为它反映的是即时效应，而不是模型中的长期净“均衡”效益。

第4节 | 空间滞后 y 和空间误差的比较

由于这里两个空间参数 ρ 和 λ 远远大于它们的标准误，因此我们可以放心地得出结论，认为数据中存在很大的空间依赖性，并且认定假设观测值间保持独立的标准 OLS 回归结果是具有误导性的。但我们会问：究竟哪个模型更好，空间滞后 y 模型还是空间相关误差模型呢？从统计上很难区分出空间滞后 y 模型和空间误差模型的好坏。这两个模型并不相互嵌套，因此也不可能将一个模型看做另一个的子集，也就是说，我们不能通过在模型中加入更多限制条件来进行假设检验。尽管可以通过正式的检验方法来比较非嵌套的模型[11]，然而，通常这些结论都是不确定的，而且很难给出一个模型优于另一个的有力证据。在这个例子中，我们可以看到两个模型的对数似然值非常相似，空间滞后 y 模型的对数似然值仅略大于空间相关误差模型。由于两个模型的参数个数相同，我们也不能说一个模型比另外一个更简约，因此，我们的经验并不能告诉我们这两个模型哪个对数据拟合得更好。一种方法是通过交叉验证或者是样本外的预测检验，但是这些方法超出了本书讨论的范围。

更重要的是，空间滞后 y 模型和空间误差模型，究竟哪一个更合适，实际上是先于理论的问题，这应当通过具体的

研究问题来考虑。如果我们期望看到——或者感兴趣的是——反馈效应，那么空间滞后 y 模型应当是一个更合适的模型。在民主的例子中，合理的预期是一个国家的民主水平会受到其他国家民主程度的影响（参见 Gleditsch，2002a；Gleditsch & Ward，2007）。相反，如果认为模型中误差的空间相关来自于系统成分中其他被忽略的特征，而国家民主水平之间没有扩散效应，这种说法就显得更不可信。因此，在这个例子中，我们相信，空间滞后 y 模型比空间误差模型更恰当。

更一般的原因是，社会科学对空间误差模型更没有兴趣。在我们看来，只有当研究者们相信误差项可以存在某些空间模式，但他们却不愿意或者是无法对误差的来源提供假设的时候，空间误差模型才更适用。这样做的原因在于，社会科学中的大多数模型在识别个体观测值特征的时候都很难抓住所有的空间聚集作用。因此，空间滞后因变量的识别问题仍然有很多工作需要做。但是如果在某个领域大部分的重要机制都已经明确并且在模型的系统部分完全识别出来，而误差项中仍然存在相关的话，这时用空间误差模型来纠正残差干扰就非常有用。总的说来，由于社会科学模型通常很少关注数据之间的依赖关系，空间误差模型可以大大改进当前的模型。

第5节 | 估计成对贸易往来中的空间性误差

为了找出空间误差模型更适用的一个例子，我们考虑它在成对贸易往来研究中的一项应用。成对(dyad)表示两个个体组成的一对，结果变量可以是某些个体特征或者个体间互动的测量，在我们的例子中表示两个国家 i 和 j 之间的贸易量。在一些情况下，我们可能希望区分 i 和 j 互动的方向，比如 $i \rightarrow j$ 表示 i 对 j 的作用。相反，没有方向的互动可以用下标定义为 $i \leftrightarrow j$。一个包含 n 个个体的系统将产生 $n \times (n-1)$ 个有方向的组对，当我们不区分往来或者互动方向时，将有 $n \times (n-1)/2$ 个无方向的组对。除了我们例子中的贸易以外，成对观测值在国际关系中非常普遍，比如当我们对估计某些特征如何影响到一个特殊事件或者行为诸如两个国家 i 和 j 之间的冲突发生的可能性感兴趣的时候。

国际关系中成对分析的传统方法是将个体互动看作这一对或者这两个个体特征的函数，同时在考虑到相关解释因素之后将两个个体视为互相独立。而空间误差模型则有利于我们处理成对观测值之间可能存在的依赖关系[12]。

国际贸易的黄金标准模型一直以来都没有发生过大的变动：这类似于牛顿的万有引力模型。贸易被看做交易国家

经济规模的函数，但是和国家之间的“距离”成反比。当前的经验研究表明很多因素都可能影响国家 i 和 j 之间的贸易程度。经验研究中最常用的贸易模型也被称之为贸易的万有引力模型，它假定两个国家之间的贸易量 $T_{i\mapsto j}$ 和它们的经济（GDP_i 和 GDP_j）、人口（P_i 和 P_j）的地理距离（$D_{i\leftrightarrow j}$）乘积成比例。该模型一般在取对数后表示成相加的形式：

$$\log(T_{i\mapsto j}) = \alpha + \beta_1 \ln(GDP_i) + \beta_2 \ln(GDP_j) + \beta_3 \ln(P_i) + \beta_4 \ln(P_j) + \beta_5 \ln(D_{i\leftrightarrow j}) + \epsilon$$

这里各种量的系数（β_1，…，β_4）应该为正值，而距离的系数（β_5）为负值。芬斯特拉、罗斯和马库森（Feenstra、Rose & Markusen，2001）以及罗斯（Rose，2004）在最近的研究中都提供了相应的例子。

万有引力模型的核心并没有涉及政治内容，但是很多社会科学家都感兴趣的是，政治因素如何影响了贸易往来。比如波林斯（Pollins，1989a、1989b）认为，政治关系可能会对贸易量产生强烈的影响，因为一个国家不太可能和与它政治关系不好的国家有很高的贸易量，或许因为商人们担心贸易受到政治因素的破坏，或许因为政府对敌对的国家的贸易采取相应的限制。莫罗、西沃森和塔瓦雷斯（Morrow、Siverson & Tabares，1998）认为民主国家更可能和民主国家进行贸易往来，同时，与其他国家相比，它们之间的军事冲突会更少影响到贸易。这些经验分析都表明这些特征会影响贸易往来。

在贸易研究中一个常常被忽略的问题是，成对的观测值之间可能并不是相互独立的。尽管有很多研究考虑到这个

问题，一对有顺序的观测值在时间先后上可能并不是独立的(Beck & Katz, 1996)，但是大部分研究都假设不同对的观测值在同一个时间点上可以被认为相互独立。然而，在贸易往来研究中，我们有足够多的理由相信这种假设不成立。因为每个国家将和众多国家组成不同的组对，所以成对数据的结果会很复杂。首先，$T_{i\to j}$和$T_{i\to k}$的贸易流通并不能看做相互独立，因为它们的贸易输出方相同。第二，通常从国家i到j的贸易流通($T_{i\to j}$)将和反方向的从j到i的贸易流通($T_{j\to i}$)呈正向关系。这种数据中还常常发现高阶依赖关系[13]。经济学家们通常会将$T_{i\to j}$和$T_{j\to i}$的值取平均，然后用分解方法分析这个三角矩阵，但这种方法却使得观测值之间的依赖关系更强。此外，众所周知，大部分报告中的贸易数据是基于其他贸易流通数据的插补估计方法得到的(例如 Rozanski & Yeats, 1994)。这种插补方法可能导致数字之间的序列相关。比如，世界银行报告的贸易数据就同班佛定律(Benford's Law)预计的首位数字分布存在显著差异，班佛定律是一种常用于检验数据质量和识别数字是否捏造的方法[14]。

该例子就适于使用空间滞后误差模型，因为我们认为某些成队组的误差项之间相互关联，而不是观测到的在贸易流通上相互关联。净流通量将取决于成对的国家数量，但仅仅依靠这一点还没有考虑到由于成对依赖性所导致的误差变化。前面我们讨论了两个个体之间地理上的距离和连接状态。在这里，我们将依赖结构定义为成对国家间拥有一个共同的成员国，但这种依赖结构并不是传统意义上的“空间”。不过这并不阻碍我们将空间概念应用到非地理距离概念中。

在这个例子中，我们的加权方案是，如果成对国家中包含 i 或 j 中的任何一个，则被看做和 $i \to j$ 这对国家相连。有关“距离”替代概念的更多讨论，可以参见贝克、格里蒂奇和比尔兹利(Beck、Gleditsch & Beardsley, 2006)，迪沃斯和伊萨德(Deutsch & Isard, 1961)，以及洛夫达哈(Lofdahl, 2002)的研究。

在经验应用举例中，我们参考格里蒂奇(Gleditsch, 2002b)研究中使用过的欧洲和非洲成对贸易数据。具体来说，我们用 $T_{i \to j}$ 表示国家 i 到 j 的输出量。非洲和欧洲的样本为我们提供了很有趣的比较，包括数据质量的变化，我们预计欧洲的贸易数据将比非洲的数据更精确，因为它们的基础设施以及经济活动监测能力都存在差异。该例子中所有的数据来源于 1998 年。在我们的样本中，贸易流动的“观察”数据来自于国际货币基金组织及其他国际机构，它们占欧洲所有成对数据的 75%左右(比如格里蒂奇 2002b 数据中原始编码以 0 或 2 开头的数据)。然而，对非洲而言，利用官方报告的数字我们将仅仅得到贸易流通数据中 15%的成对数据。在该例子中，对于欧洲我们将仅仅使用官方报告的数据，而在非洲贸易流通分析中，我们将使用所有可能的、甚至是有争议的数据来源作为估计。

标准的“万有引力模型”的变量包括经济规模，两个成员国的人口(数据来源于 Gleditsch, 2002b)和它们首都之间的距离。此外，我们的模型参考了现有文献中贸易的政治决定因素，并包括两个国家政治取向的相似性，这种相似性的测量是通过两个国家在联合国投票记录中的 S 相似性得分得到的(见 Gartzke, 1998; Signorino & Ritter, 1999)。民主的

测量来自于POLITY Ⅳ数据。我们对数据进行适当的调整，包括对没有被纳入到自由屋(Freedom House)原始数据中的国家的POLITY数据也进行了估计[15]。根据贾格斯和格尔(Jaggers & Gurr, 1995)对制度化民主测量的21点量表，我们选取了两个数值中较低的那个，经过重新调节比例，使得所有的值都为正。最后，我们也考虑了成对的两个国家是否曾卷入军事化国际争端(见Jones、Bremer & Singer, 1996)。

```
source("chapter3data.R")
tab3.sem <- errorsarlm(logtrade ~ logdem + logapop + logbpop +
      logargdppc + logbrgdppc + logs + logdist + logmid,
      data = logdat98, na.action = na.omit,
      nb2listw(dlist, style = "W"), method = "eigen")
summary(tab3.sem)
logLik(tab3.sem)
```

表3.2和表3.3分别列出了欧洲和非洲成对贸易往来的OLS和空间相关误差模型估计结果。从$\hat{\lambda}$可以看出，非洲和欧洲样本中的成对贸易伙伴之间都存在很强的空间正向相关。同时，通过比较OLS和空间误差模型估计结果可以看出，当我们考虑到成对成员的残差空间相关，而不是把它们视为相互独立的观测值时，原有文献中强调的贸易变化的政治决定因素发生了很大的改变。尤其是当我们考虑到成对成员的空间相关关系之后，MID所起的负向作用的估计系数在欧洲样本中降低了大约25%，在非洲样本中降低了大约40%。在非洲样本中，民主的估计系数降到了原有大小的

表 3.2 出口,欧洲：$T_{i\to j}$

变量	OLS			SEM		
	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距	−32.70	0.67	−48.82	−33.94	1.71	−19.90
民主对数	0.38	0.06	5.93	0.43	0.10	4.38
i 人口对数	0.86	0.02	40.37	0.89	0.03	31.46
j 人口对数	0.75	0.02	34.93	0.77	0.03	27.33
i 人均 GDP 对数	1.54	0.04	35.23	1.56	0.06	17.35
j 人均 GDP 对数	1.01	0.04	23.07	1.03	0.06	7.66
S 对数	0.33	0.05	6.92	0.35	0.05	7.69
$i\leftrightarrow j$ 距离对数	−0.34	0.01	−24.33	−0.34	0.01	−25.83
$i\leftrightarrow j$ 争端对数	−1.94	0.27	−7.14	−1.48	0.29	−5.01
$\hat{\lambda}$				0.98	0.01	73.73
N		1500			1500	
自由度(df)		8			9	
对数似然值 (Log likelihood)		−2324.8			−2239.668	

注：SEM 表示空间误差模型。

表 3.3 出口，非洲：$T_{i \to j}$

变量	OLS			SEM		
	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距	−7.41	0.33	−22.38	−7.47	1.45	−5.16
民主对数	−0.04	0.04	−1.08	−0.01	0.05	−0.15
i 人口对数	0.26	0.01	20.51	0.26	0.02	14.45
j 人口对数	0.23	0.01	17.81	0.23	0.02	12.55
i 人均 GDP 对数	0.38	0.02	17.96	0.38	0.03	12.78
j 人均 GDP 对数	0.31	0.02	14.82	0.31	0.03	10.55
S 对数	3.41	0.40	8.50	3.43	0.47	7.24
$i \leftrightarrow j$ 距离对数	−0.17	0.01	−20.81	−0.17	0.01	−22.21
$i \leftrightarrow j$ 争端对数	−0.71	0.18	−3.85	−0.42	0.18	−2.37
$\hat{\lambda}$				0.99	0.01	124.2
N		2550			2550	
自由度(df)		8			9	
对数似然值 (Log likelihood)		−3096.2			−2945.9	

1/4，而在欧洲样本中该系数增大了约15%。此外，空间误差模型中的个体系数估计值的标准误一般比OLS结果更大，这表明如果将个体成对观测值视为相互独立，由于模型中错误的标准误可能导致对估计结果过于肯定。更一般地说，尽管按照惯例标准，系数中并没有从“显著”变为“不显著”的情况发生，但很多独立成对个体假设下得到的明显结果，在考虑到空间依赖关系之后，变得不那么稳健。

第6节 | 小结

在本章中,我们介绍了有关空间依赖关系的空间相关误差模型。在通常情况下,由于我们很难(尽可能地)完全基于统计标准判断空间相关误差模型和空间滞后 y 模型孰好孰坏,所以研究者们应当考虑这两种模型哪种能够为空间依赖关系提供最可信的解释。我们前面已经提到,空间滞后 y 模型更适合于处理当因变量的邻近值的变化对该个体因变量产生直接影响的情况,而空间相关误差模型更适用于当我们确信模型系统部分某些未观测到的特征,可能导致模型的误差出现空间相关模式的情况。在成对个体相互依赖的例子中,某个国家与很多观测值构成了不同的组对。这也表明,空间依赖概念可以从地理距离扩展到度量学的距离。

第4章

扩　展

本书前面的章节讲解了将空间类型引入社会科学数据中进行分析的必要性和益处。我们介绍了如何将这种方法纳入常见的线性回归框架——空间滞后因变量模型和空间误差模型。在空间滞后因变量模型中,与 y_i 相连的单位将对 y_i 产生影响;在空间误差模型中,相连观测值的误差之间存在空间相关。这两种是最广泛使用的空间回归模型,它们有很多广泛应用。然而,空间回归模型也有很多其他的类型,以及在很多我们没有提及的情况下的扩展应用,并且我们前面的关注也仅局限在连续变量的横截面数据上。在本节中,我们列出了一些空间回归模型的扩展应用,以及空间分析可能面临的棘手问题。尽管我们的回顾非常简略并且不能提供涵盖这些扩展和替代方法的实际例子,但是我们提供了进一步阅读的参考建议。比万德、佩勃斯玛和戈麦斯—卢比奥(Bivand、Pebesma & Gomez-Rubio, forthcoming)提供了这些方法基于 R 统计软件的教学材料。

第 1 节 | 识别连接性

如何建构和处理观测值之间的连接,是分析中研究者所面临的一个关键问题。大多数空间回归模型的应用都事先假定了观测值之间的连接图。如何建立这些连接取决于有关观测值在实际中如何关联的理论或者直觉。在实际操作中,这些建立方法都是出于方便考虑或者是基于最新的常用方法。研究者们需要注意的是,选择不同的连接方法和编码方法可能意味着认识世界的不同观点。不同的结果可能造成个体直接连接形式的差异,但这并不是什么大问题。更微妙的地方在于这些选择也会影响空间结构中的空间乘子和模型中的协方差结构(Wall, 2004)。即便是在地理距离的连接情况下,同样的空间布局可能由于研究者不同的决定而产生不同的连接结构。为了说明这一点,我们可以参考三种常见空间编码的差异:车(Rook,共同边界)、象(Bishop,共同顶点)和后(Queen,同时包括边界和顶点),见图 4.1 所示的美国局部地图。科罗拉多州和犹他州是具有共同边界和共同顶点的邻州。科罗拉多州和亚利桑那州没有共同的边界,但有共同的“顶点”。在当前的世界国家政治边界地图中,我们只发现了一个类似的情况:非洲西南的卡普利维地带(the Caprivi Strip)。



图 4.1 美国四角区域

更典型的情况是：如果个体距离在某个范围内，也就是行政中心或者地理质心或中心之间的最短距离，研究者们就将它们视为“相近”。格里蒂奇和沃德（Gleditsch & Ward, 2001）讨论了常用的中点测量存在的一些问题，也就是当个体的行政中心和边界相差很远或者是个体的奇怪形状导致中心不在版图之内的情况。划定过于窄的范围可能生成很多岛屿，这个问题我们在第1章新西兰的例子中提到过。从澳大利亚阿利斯斯普林斯（Alice Springs）到新西兰克赖斯特彻奇（Christchurch）大约为4100千米，相当于巴黎到达累斯萨拉姆（Dar es Salaam）的距离。这表明如果我们用澳大利亚和新西兰版图中心的距离作为标准并用到其他国家的话，那么大部分非洲国家、中东和亚洲国家都将成为法国的邻国。划定标准过大的话将导致所有个体都相连。两幅国家连接图中（图4.2），当中心连接距离从400千米（图a）变为4000千米（图b）时，连接密度急剧增加。如果针对特定的个体设定连接规则，或者选取 k 个最相邻的观测值相互连接，则会引出另外一个问题，即为什么不对其他的个体也使用同

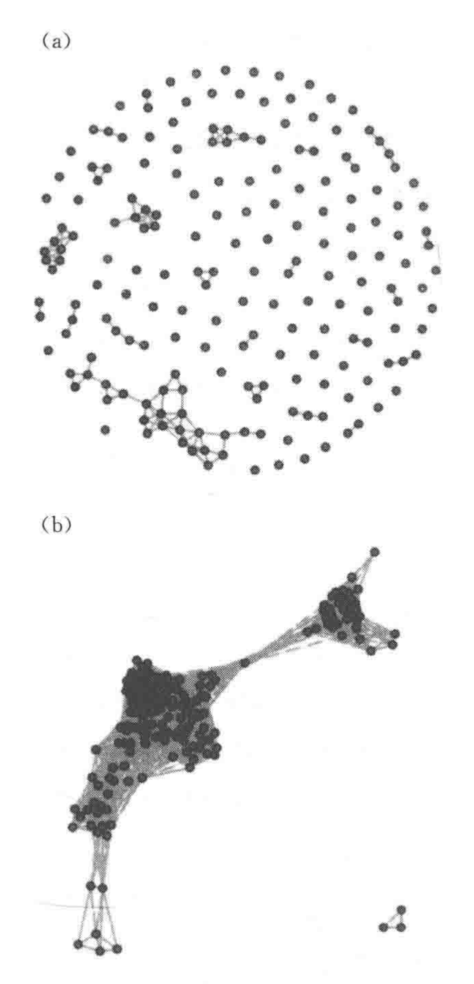


图 4.2 国家之间的连接,4000 千米和 400 千米的距离,基于版图中心间的距离:(a)400 千米;(b)4000 千米

样的标准。然而,这种特定规则在应用研究中不仅可能有用,而且是必要的。总的说来,对连接编码的不同选择总会对经验结果产生实质性的影响,因为通过不同的网络扩散将带来不同的结论。

基于非地理性测量的连接,比如贸易往来,可能存在其

他的问题。更确切地说,如果非地理性的距离测量是基于实际空间回归模型中已有的变量,那连接将可能不是外生性的,这将导致模型识别和估计上的问题。研究者需要设计出与空间互动过程相匹配的连接矩阵。尽管拟合优度和交叉验证方法有助于剔除错误选择,但是连接性的设计其实是一个理论问题,并不可能使用简单的诊断或者探索方法来定义唯一"正确"的连接方式。我们还要强调的是,连接性识别上的困难也增大了检验空间依赖关系的零假设的难度,因为拒绝零假设只是针对某种特定连接方式而言的。

处理连接性

在识别出连接矩阵以后,如何在分析中处理连接矩阵则是另外一个问题。我们应当对所有的连接赋予相同的权重,还是应当根据观测值大小或重要性给予一些观测值不同的权重?在本书的例子中,我们假设俄罗斯和爱沙尼亚和与它们相连的国家之间的权重相同。然而,并没有谁规定所有的连接就应当使用相同权重。针对具体研究问题,研究者可能尝试不同的加权方法。

在回归模型中,我们仅仅考虑了行标准化矩阵 W 的情况,也就是所有的连接权重相加为 1。这种标准化的优点在于空间滞后 y^s 和 y 具有同样的潜在测量标准或单位。然而,标准化是否合理还是应当具体问题具体分析。举例来讲,默多克等人(Murdoch et al., 1997)关注过一个国家的污染排放量如何受到其他国家排放的影响。这个问题涉及总的污染排放量,这时利用相连国家的数量来对连接矩阵进行标准

化可能就不适用了。

分析者可以将空间统计文献中惯用的做法作为参考，并仔细分析它们在自己的研究中是否行得通。比较有用的方法是多考虑几种替代方法。

一个对多个的连接

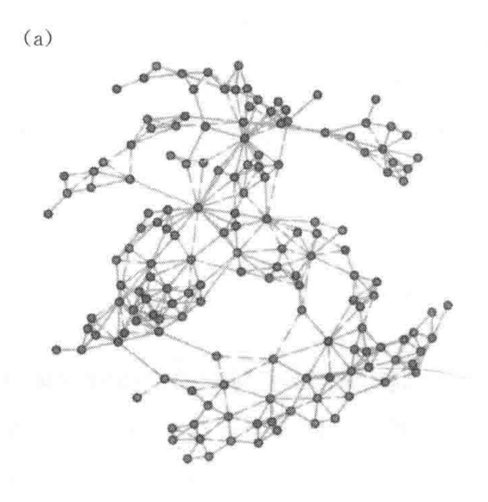
到目前为止，我们讨论了具有单一空间依赖项并表示成单一连接矩阵的例子。在很多情况下，可能会出现多种连接网络或者依赖形式的情况。通常比较可行的方法是根据地理距离或者其他政治网络（比如贸易合作、文化相似性，或者种族划分、职业划分）考虑几种不同的连接方法（见 Beck et al.，2006；Lacombe，2004；Lin et al.，2006）。直接影响不仅可能来源于一阶连接，也可能是高阶连接。图 4.3 画出了前面例子中提到的 158 个国家的一阶和高阶连接。

空间滞后 y 模型可以被推广到包括两个（或更多）不同连接矩阵 W^A 和 W^B 的形式，并通过如下表达式分别估计参数 ρ_1 和 ρ_1 各自的相对影响。

$$y_i = x_i\beta + \rho_1 w_i^A y + \rho_2 w_i^B y + \epsilon$$

扩展后的空间滞后 y 模型变得比标准空间自回归模型更难估计。假如这两个矩阵差别足够大并且不包含重复信息，那么这个模型便可以估计。如果这两个矩阵太相似，那么就会出现诸如经典回归模型中的共线性问题。前面讨论过的最大似然法也可以应用到这种情况（尽管还没有在 R 中实现）。这种模型也可以用工具性变量来估计。

(a)



(b)

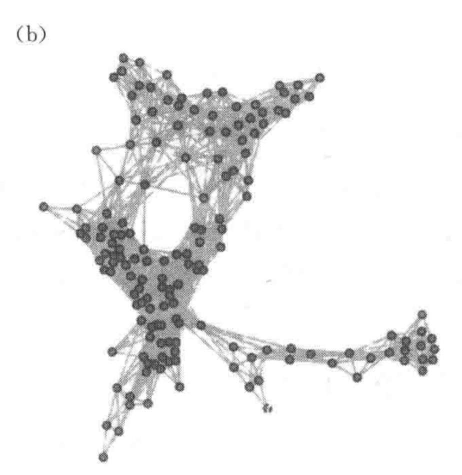


图 4.3　158 个国家一阶和二阶连接，基于最近邻国距离为 200 千米：(a)一阶；(b)一阶和二阶

第 2 节 | 推论与模型评估

和社会科学中的大多数数据一样，空间数据也不是来自于随机样本。空间分析需要一个相对完整的空间覆盖，这是因为包含太多缺失值的数据可能使对空间聚集影响或者相邻单位之间影响的统计推论变得毫无意义。虽然经典统计推论在很大程度上是基于渐进性假设，但在很多空间情况中这一点却很难证明。实质上，这要求某地区邻地的数量不会随着该地区的大小而剧烈变化。即便满足这一点，研究中的空间数据看上去并不像一个样本，而是像截面上的某个地区或者是整个世界。这使得我们需要找出符合研究背后社会过程的合理的模型，也就是原则上产生观测数据的模型。经典的方法可能是基于一种概括性的“超总体”(super-population)概念，也就是观测到的空间类型在现实中的体现，但是这种概念在贝克、韦斯特和韦斯(Berk、Western & Weiss，1995)那种所谓的“明显的总体”(apparent population)的空间分析中并不适合。

这种难题的一个可能的解决办法是以探索性的态度看待估计结果，并且按照格斯尔(Geisser，1974、1975)的方法，用空间回归估计中没有用到的数据进行交叉检验，来检查模型的估计好坏。在空间背景下，这种方法可以通过利用后一

个时间段或不同空间领域的观测数据来实现。比如，比万德(Bivand, 2002)就将数据分成两个地理区域，根据它们对数据另一半的观测值的预测能力，来评估不同建模方法的好坏。

离散和潜变量

我们前面假设 y 应当是连续变量。然而社会科学家感兴趣的很多现象都是离散的，也就是说以二分的或者计数的形式出现。另外，看到的结果也可能来自于潜在社会过程的一部分。这就好比线性回归只是这些数据的次优选择一样，这里估计的模型也不一定最适合这些数据。然而，将滞后因变量或者自回归过程扩展到二分情况或者计数数据也是可能的，比如在自相关逻辑(autologistic)模型中，相邻个体的 y 值会影响 $Pr(y_i = 1)$的情况(Besag, 1972、1974; Christensen & Waagepetersen, 2002; Huffer & Wu, 1998; Ward & Gleditsch, 2002)。估计这些模型将比估计连续变量的情况复杂得多，因为 y 同时出现在方程的两边，这使得似然值很难处理。传统的方法在估计中将相连观测值的 y 视为固定，但是当前的计算能力使得利用模拟方法估计整个似然值成为可能。

空间异质性

通常回归中的主要效应为固定效应，这表明自变量和因变量之间的关系是无处不在的。然而这种关系可能在世界

上某个地方不同于其他地方，空间异质性指的就是这种效应和地理条件有关。空间异质性既为我们提供了了解研究现象的机会，也给研究带来了困难。一方面，它让我们可以分解回归结果，使得这些结果能更好地反映不同地区的情况。另一方面，它又违背了标准回归中所有分析数据方差相同的假设。地理加权回归，即GWR(Geographically Weighted Regression)，作为一种数据探索的开创性方法，让我们利用某一位置的相邻观测值作为权重，估计每一个地理位置的回归系数。布伦森、弗泽林哈姆和查尔顿(Brundson、Fotheringham & Charlton, 1996)发展了空间分析中的这种方法，更多详细内容可以见弗泽林哈姆、查尔顿和布伦森(Fotheringham、Charlton & Brundson, 2002)的研究。在政治科学中，一个最近的例子来自卡尔沃和埃斯科拉(Calvo & Escolar, 2003)；人口学中一项有意思的应用是厄舍克和皮纳斯古鲁(Işik & Pinarcioğlu, 2007)的研究。

点和地理统计数据

前面的方法都是将地理空间看做可以划分的。比如国家被看做一个个格子，表明每个国家都可以在地图上找到一个方格，没有哪个国家占有超过一个方格的位置。在很多数据中，这种方法都是有用的，但并不是所有的现象都可以被视为区域或者格子，通常数据也不是按照这种格式来表现的。事实上，很多类型的数据都是地理上点数据的形式，这样每个观测值确切或者近似的位置是一个连续的地质结构，而不是想象中的一个格子。地理统计学方法试图根据某个地

理区域具体位置的信息创建空间共变模型，从而使连续的地理学变成地理统计形式。一种方法称为克里金(Kriging[16])法，由马特隆(Mtheron, 1963)正式发展而来，但以南非采矿工程师丹尼·克里格(Danie G. Krige)的名字命名，因为他开创了距离加权后测绘平均黄金等级的方法。这种方法广泛用于地球物理科学，现在也用于社会科学(Cho & Gimpel, 2007)。尽管以往研究用到的都是大样本汇总层面的数据或者没有空间识别特征的数据，但是现在可用的地理细分数据或者明显与地理相关的数据变得越来越多。

多层模型

在贝斯格(Besag, 1974)的早期贡献之后，有相当多的研究讨论了条件自回归模型(Conditionally Autoregressive Model, CAR)。在条件模型中，在某个地点观测到的随机变量取决于其邻近的外生观测值。在多变量和多层模型中，不仅空间滞后变量需要是外生的，而且其他的解释变量也需要为外生。当前很多研究都在利用这种方法，有的是关注几个结果因变量。当前这些工作可以见金、班纳吉和卡林(Jin、Banerjee & Carlin, 2007)以及茹和海尔德的研究(Rue & Held, 2005)。

另一种建立空间变化模型的相关方法是用多层方法考察本地变化的来源。多层空间模型是将不同分析层次的不确定性来源整合在一起。这种模型通过概率分布将不同层次的分析联系起来。在民主与发展的例子中，可以包括如下层次：(1)本国国内政党派系和机构的变化，它们会影响

日常政治和经济的波动;(2)邻里效应,指某个相邻国家和另一国家之间存在很强的联系并且受其影响;(3)基于一系列国家所产生的区域效应,包括有些沿着地区边界运作的组织[1];(4)全球化的力量,它会不同程度地影响到每一个国家,比如全球市场上的某些商品。模型中如果明确和详细说明了这些变化的来源,则属于多层模型。

当前处理这种观点的方法都是基于贝叶斯方法,也就是它依赖于迭代法[马尔可夫链蒙特卡罗方法(Markov chain Monte Carlo)、吉布斯抽样(Gibbs sampling)、Metropolis-Hasting 等算法]来得到一套所有层次的空间过程的参数分布。这种方法需要很多计算,但是很有前景。当前的 R 软件包 spBayes 也可以帮助实现单变量和多变量空间模型的马尔可夫链蒙特卡罗计算(Finley、Banerjee & Carlin, 2007)。沃勒、卡林、夏和盖尔芬德(Waller、Carlin、Xia & Gelfand, 1997)提供了一项很有影响的应用,班纳吉等人(Banerjee et al., 2004)为多层方法提供了很好的综述。

时间序列数据

我们已经讨论了在同一时间段里横截面观测数据的模型估计。社会科学的很多分析都是基于时间序列的横截面(Time Series Cross Section, TSCS)数据结构,也就是同一个个体有多个不同时间点上的观测。空间滞后 y 模型也可以扩展到 TSCS 数据的情况:

---

[1] 比如前文提到的经济合作和发展组织(OECD)就是这样一种情况。该组织中的成员国之间彼此的相互影响可能大于来自其他国家的影响。——译者注

$$y_{i,t} = x_{i,t}\beta + \rho w_i y_{i,t} + \epsilon_{i,t}$$

该模型可能会面临时间上的序列相关问题，因为 $y_{i,t}$ 可能会和 $y_{i,t-1}$ 非常相似，进而造成误差独立假设的问题。一种解决方法是通过加入 y 的时间间隔，从而有：

$$y_{i,t} = \phi y_{i,t-1} + x_{i,t}\beta + \rho w_i y_{i,t} + \epsilon_{i,t}$$

如果我们想要同时说明时间和空间的依赖性，估计包含联立空间依赖性的 TSCS 模型是相当困难的。如果我们在右边加入滞后因变量，y 的误差项 ϵ 的雅克比行列式转换（Jacobian of the transformation）将变得相当复杂，并且据我们所知，当前还没有人得到这种模型的满意估计。然而，如果假设 $y_{i,t}$ 对相邻的 y 的影响的发生具有一个时间滞后（比如 $y_{i,t-1}$）的话，就可以用 OLS 方法，因为与 y 相关的邻近值可以被视为在时间 t 之前就预先确定。这将得到：

$$y_{i,t} = \phi y_{i,t-1} + x_{i,t}\beta + \rho w_i y_{i,t-1} + \epsilon_{i,t}$$

空间效应中引入时间滞后量通常被认为和假设瞬时效应一样都是合理的。此外，还可以通过对模型估计残差进行适当的检验，尤其是交叉检验和样本外的探索方法，来检验模型在多大程度上成功解释了空间和时间依赖性（更多讨论，见 Beck et al.，2006）。

第 3 节 | 小结

空间依赖关系在很多社会现象中都发挥着重要作用。将空间层面引入分析中也是完全可行的，但是需要附加一些假设和信息。随着统计和计算机技术的发展，空间数据分析的障碍得以消减，我们期待它能给社会科学家们感兴趣的社会和空间过程带来新的见解。以往的经验让我们相信社会科学数据中具有很多未发现的依赖特征。只要将这些特征中的一部分考虑进来就可以产生全新的重要的启发。



附　录

附录 | 软件选项

在很长一段时间里，标准统计软件包都不能提供空间估计值，这使得感兴趣的研究者需要自己编程序或者购买安瑟林的 *SpaceStat* 软件。在过去几年中这种情况发生了很大的改变。在本部分中，我们将介绍一些可用的软件选择。

很多软件选项都依赖于奥德方法(Ord approach)，在最优化之前估计矩阵 W 的特征值，不过现在一些其他方法也用更快的佩斯和巴里方法(Pace & Barry approach)。很多软件包都要求输入 $n \times n$ 的全秩矩阵。但这对于大样本的数据集来说很难做到。因为通常连接矩阵中会有很多 0 存在，软件选项中如果有用稀疏矩阵(sparse matrix)表示的话，就可以应用到更大的数据集上。

这里我们列出空间分析的一些软件选项。

1. 安瑟林 *SpaceStat* 软件如今已不是安瑟林个人掌管的一个单独的软件包，而是卖给了一个商业公司，成为地理视觉化程序 *TerraSeer* 的一部分(见http://www.terraseer.com/products_spacestat.php)。该软件的费用非常昂贵，即便对学术用途而言也是如此。旧版本的 *SpaceStat* 在 MS-DOS 系统中运行，使用传统的菜单式界面。它也是依赖于奥德方法来计算加权矩阵的行列式，同时需要全部的矩阵表达

式以进行估计。我们对当前的 *TerraSeer* 产品不太了解，因此不知它和以前的版本有何差别。

2. 安瑟林和他的同事们开发出一种新的软件包叫做 GeoDa，可见 http://www.geoda.uiuc.edu/。GeoDa 可以进行探索性空间数据分析和简单的空间回归分析。GeoDa 界面完全是通过点击来完成的，而不需要任何编程；然而它不允许用户自定义或者修改其中的设置，这和一般化的统计软件包不同。安瑟林（Anselin、Syabri & Kho，2004）等人表明，这种软件主要是作为初学软件包，当用户了解了 DeoDa 中的技术之后，可以逐步过渡到 R。

3. 皮萨蒂(Pisati)编写的 *Stata* 宏 spatreg 可以估计空间自回归和误差模型。这个程序或者宏不好的地方在于它依赖于奥德方法，因而需要全部的矩阵表达式。*Stata* 的标准版本（standard Intercooled version)里同样还对矩阵大小有限制。参见 *Stata* 的技术报告（technical Bulletin）sg162。安装帮助见 *Stata* 中的 help stb。

4. 比万德(Roger Bivand)开发了本书中讨论的模型的 R 软件包(*spdep*)。这个软件包还允许连接矩阵具有稀疏形式（sparse list)。比万德还开发了各种整合 R 和 GRASS 的材料，GRASS 是一种开源的 GIS 程序。此外，很多功能还可以用来制作地图，以及从 R 的 *Arcview* 的形文件（shapefiles)中提取信息。更多有关该软件包的细节可见：http://cran.r-project.org/src/contrib/Descriptions/spdep.html。这些功能和软件后台都是开放源代码并且免费的。

5. 一些有关空间分析的 *MATLAB* 教材也已出版。佩斯和巴里的空间统计学（Spatial Statistics）可以在 http://

www. spatial-statistics. com/免费获得。MATLAB 本身并不是免费的。勒萨热(LeSage)的空间计量经济学(Spatial Econometrics)工具箱,可以在 http://www. spatial-econometrics. com 获取,它在估计大样本的数据集时非常有用,并且可以用于利用 saw()命令,估计具有两个连接矩阵的空间自回归模型。

6. 来自 ESRI 公司最新版本的商业软件包 *ARCINFO* 包括很多可以用于空间形式数据集的统计分析工具,尤其是其中的 Statistical Analyst 工具包。它特别擅长计算邻里数据和进行分类分析。

7. Splus 来自于 Insightful 公司。和 R 一样,它也是基于 S 统计语言。它包括一个能提供很多空间相关数据分析的模块(SpatialStats)。其他工具还包括地理统计的、点状的以及格子状的空间数据。

8. WINBUGS (http://www. mrc-bsu. cam. ac. uk/bugs/)和 GeoBUGS 是两个针对贝叶斯分析的程序。GeoBUGS 由一个流行病学家开发来作为 WinBUGS 的附加程序。它支持(相对较小的)空间模型的贝叶斯分析。

9. 斯卡本伯格和果特威(Schabenberger & Gotway, 2005)提供了 SAS 中分析空间数据的宏和程序的扩展集,这些可以从出版商的网页上找到:www. crcpress. com。

10. 空间多层方法可以很容易在 R 软件包 spBayes 中找到,它可以进行常用的 MCMC(Markov Chain Monte Carlo)计算(Finley et al., 2007)。

注释

[1] 然而在社会科学中很少应用点数据。一个例外来自于最近曹和金佩尔(Cho & Gimpel, 2007)的研究。

[2] 葛兰德(Grenander, 1954)指出,即便对于均值的最小无偏估计也不应该忽略相关联的观测值的值。

$$\hat{\mu}=\frac{\left[y_1+(1-\rho)\sum_{i=2}^{n-1}y_i+y_n\right]}{[n-(n-2)\rho]}$$

[3] 在空间点过程条件下,有时这被称为合算统计量(Join-count statistics),因为它们计算了包含相似联合分布的邻近点的数量。

[4] 吉尔里(Geary)和莫兰(Moran)在很多领域都作出了重要贡献。吉尔里因斯通—吉尔里效用函数(Stone-Geary utility function)和计算真实收入的等同购买力的国际比较而出名。

[5] 数据获取方式 http://privatewww. essex. ac. uk/~ksg/mindist. html。

[6] 莫兰 I 用于处理单个变量。当模型中明显存在多个变量时,比如 OLS 残差的例子,我们推荐使用一个略微修改的莫兰统计量,用于防止过高估计空间相关性。实际上,由于差别细微,通常在两种情况下都使用标准的莫兰统计量。特菲尔斯多夫(Tiefelsdorf, 1972)发展出一种潜在分布的鞍点(saddlepoint)近似法,这种方法在匹配分布尾部的很多情况下都比莫兰 I 要好。在这个问题上我们感谢罗杰·比万德(Roger Bivand)的建议。R 的 spdep 函数 lm. morantest. sad()可以执行特菲尔斯多夫的鞍点法。一个更直观的方法是利用拉格朗日乘子检验(Lagrange multiplier test)来检验空间自相关的具体形式,在下一章当中我们将会详细叙述。

[7] 正如海宁(Haining, 2003:276—283)提出的,如果自变量之间存在很强的空间相关性,标准莫兰 I 可能过高估计残差的空间相关性。

[8] 用巴尔多夫—尼尔森(Barndorff-Nielsen)鞍点调整方法得到略微减小的估计:相对应的值为 6. 9。

[9] 还可以设想出更多类似的分解方法,比如引入多层协变量使得它们仅仅影响到某个具体区域或者行政区内的观测值。本书中不讲解这种方法。

[10] 2004 年选举数据可以从 http://www. fec. gov/pubrec/fe2004/federalelections2004. pdf 获取。

[11] 有关非嵌套检验的例子可以参见 Clarke(2001)。

[12] 对该种以及其他依赖关系更详尽的考查可以参考 Ward & Hoff (2007)；针对二元因变量，见 Ward、Siverson & Cao(2007)。

[13] Wasserman and Faust(1994)总结过这种成对数据中三个一组的情况，主要是社会网络分析。

[14] 首位数字法则(first-digit law)，以物理学家弗兰克·班佛命名，它说明数据中的第一位数字最常见为 1，越大的数字将越少见，或者更确切地说，数字出现的频数接近于。这种法则适用于大样本分布的、自然产生的数据，同时它还表明如果数据的分布和这种分布相差很大，则表明数据质量较低甚至可能是捏造的数据。

[15] 见 http://privatewww.essex.ac.uk/~ksg/polity-data.html。

[16] 根据发明者的名字，Kriging 应当被读做“Kricking”。

参考文献

Adolph, C. A. (2004). *The dilemma of discretion: Career ambitions and the politics ofcentral banking*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht, The Netherlands: Kluwer.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis*, *27*, 93—115.

Anselin, L., Syabri, I., & Kho, Y. (2004). *GeoDa: An introduction to spatial data analysis (Typescript)*. Urbana-Champaign: Department of Agricultural and Consumer Economics, University of Illinois.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: Chapman & Hall.

Baybeck, B., & Huckfeldt, R. (2002). Urban contexts, spatially dispersed networks, and the diffusion of political information. *Political Geography*, *21*, 195—220.

Beck, N., Gleditsch, K. S., & Beardsley, K. (2006). Space is more than geography: Using spatial econometrics in the study of political economy. *International Studies Quarterly*, *50*, 27—44.

Beck, N., & Katz, J. N. (1996). Nuisance vs. substance: Specifying and estimating timeseries—Cross-section models. *Political Analysis*, *6*, 1—36.

Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations (with discussion). *Sociological Methodology*, *25*, 421—485.

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B, Methodological*, *34*, 75—83.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, *36*, 192—225.

Bivand, R. (2002). Spatial econometrics functions in R: Classes and methods. *Journal of Geographical Systems*, *4*, 405—421.

Bivand, R., Pebesma, E., & Gomez-Rubio, V. (forthcoming). *Applied spatial data analysis with R*. New York: Springer.

Boots, B. N., & Tiefelsdorf, M. (2000). Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems*, *2*, 319—348.

Brundson, C., Fotheringham, A. S., & Charlton, M. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, *28*, 281—298.

Burkhart, R., & Lewis-Beck, M. (1994). Comparative democracy: The economic development thesis. *American Political Science Review*, *88*, 903—910.

Calvo, E., & Escolar, M. (2003). The local voter: A geographically weighted approach to ecological inference. *American Journal of Political Science*, *47*, 189—204.

Cho, W. K. T., & Gimpel, J. G. (2007). Prospecting for (campaign) gold. *American Journal of Political Science*, *51*, 255—268.

Christensen, O. F., & Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, *58*, 280—286.

Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science*, *45*, 724—744.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cliff, A. D., & Ord, J. K. (1971). Evaluating the percentage points of a spatial all tocorrelation coefficient. *Geographical Analysis*, *4*, 51—62.

Cressie, N. A. C. (1993). *Statistics for spatial data* (rev. ed.). New York: Wiley.

Dalgaard, P. (2002). *Introductory statistics with R*. Berlin: Springer.

Deutsch, K. W., & Isard, W. (1961). A note on a generalized concept of effective distance. *Behavioral Science*, *6*, 308—311.

Feenstra, R. C., Rose, A. K., & Markusen, J. R. (2001). Using the gravity model to differentiate among alternative theories of trade. *Canadian Journal of Economics*, *34*, 430—447.

Finley, A. O., Banerjee, S., & Carlin, B. P. (2007, April). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, *19*(4). Retrieved Oc-

tober 29, 2007, from http://www.jstatsoft.org/v19.

Fotheringham, A. S., Charlton, M., & Brundson, C. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. New York: Wiley.

Franzese. R. (1999). Partially independent central banks, politically responsive governments, and inflation. *American Journal of Political Science*, *43*, 681—706.

Franzese, R., & Hayes, J. C. (2007). Spatial econometric models for the analysis of TSCS data in political science. *Political Analysis*, *15*, 140—164.

Gartzke, E. (1998). Kant we all just get along? Opportunity, willingness and the origins of the democratic peace. *American Journal of Political Science*, *42*, 1—27.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, *61*, 101—107.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*, 320—328.

Getis, A., & Boots, B. (1978). *Models of spatial processes*. Cambridge, UK: Cambridge University Press.

Getis, A., & Ord, J. K. (1996). Local spatial statistics: An overview. In P. Longley & M. Batty (Eds.), *Spatial analysis: Modelling in a GIS environment* (pp. 261—277). Cambridge, UK: Geoinformation International.

Geyer, C. J., & Thompson, E. A. (1992). Constrained Monte Carlo, maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *Methodological*, *54*, 657—699.

Gleditsch, K. S. (2002a). *All international politics is local: The diffusion of conflict, integration, and democratization*. Ann Arbor: University of Michigan Press.

Gleditsch, K. S. (2002b). Expanded trade and GDP data. *Journal of Conflict Resolution*, *46*, 712—724.

Gleditsch, K. S., & Ward, M. D. (1997). Double take: A re-examination of democracy and autocracy in modern polities. *Journal of Conflict Resolution*, *41*, 361—382.

Gleditsch, K. S., & Ward, M. D. (2000). War and peace in time and space: The role of democratization. *International Studies Quarterly*, *44*, 1—29.

Gleditsch, K. S., & Ward, M. D. (2001). Measuring space: A minimum distance database and applications to international studies. *Journal of Peace Research*, *38*, 749—768.

Gleditsch, K. S., & Ward, M. D. (2007). The diffusion of democracy and the international context of democratization. *International Organization*, *60*, 911—933.

Grenander, U. (1954). On the estimation of regression coefficients in the case of autocorrelated disturbance. *Annals of Mathematical Statistics*, *25*, 252—272.

Griffith, D. A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In S. Arlinghaus (Ed.), *Practical handbook of spatial statistics* (pp. 65—83). Boca Raton, FL: CRC Press.

Griffith, D. A. (2003). Using estimated missing spatial data with the 2-median model. *Annals of Operations Research*, *122*, 233—247.

Haining, R. (2003). *Spatial data analysis: Theory and practice* (1st ed.). Cambridge, UK: Cambridge University Press.

Holmes, T. J. (2006, February). *Geographic spillover and unionism*. National Bureau of Economic Research (Working Paper Series 12025). Retrieved October 17, 2007, from http://www.nber.org/papers/w 12025.

Hubert, L. J., Golledge, R. G., & Constanzo, C. M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, *12*, 224—233.

Huffer, F. W., & Wu, H. (1998). Markov chain Monte Carlo for autologistic, regression models with application to the distribution of plant species. *Biometrics*, *54*, 509.

Imai, K. (2005). Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review*, *99*, 283—300.

Işik, O., & Pinarcioğlu, M. M. (2007). Geographies of a silent transition: A geographically weighted regression approach to regional fertility differences in Turkey. *European Journal of Population*, *22*, 399—421.

Jaggers, K., & Gurr, T. R. (1995). Tracking democracy's "Third Wave" with the Polity III data. *Journal of Peace Research*, *32*, 469—482.

Jin, X., Banerjee, S., & Carlin, B. P. (2007). Order-free coregionalized

areal data models with application to multiple disease mapping. *Journal of the Royal Statistical Society, Series B*, *69*, 817—838.

Johnson, S. (2006). *The ghost map*. New York: Riverhead Books.

Jones, D. M., Bremer, S. A., & Singer, J. D. (1996). Militarized interstate disputes, 1816—1992: Rationale, coding rules, and empirical applications. *Conflict Management and Peace Science*, *15*, 163—213.

Keele, L., & Kelly, N. J. (2006). Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political Analysis*, *14*, 186—205.

Kenny, D. (1981). Interpersonal perception: A multivariate round robin analysis. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences: A volume in honor of Donald T. Campbell* (pp. 288—309). San Francisco: Jossey-Bass.

Kidron, M. (1981). *The state of the world atlas*. New York: Simon & Schuster.

Lacombe, D. (2004). Does econometric methodology matter? An analysis of public policy using spatial econometric techniques. *Geographical Analysis*, *36*, 105—118.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: Wiley.

Lee, C.-S. (2005). Income inequality, democracy, and public sector size. *American Sociological Review*, *70*, 158—181.

Leontief, W. W. (1986). *Input-output economics*. New York: Oxford University Press.

Lin, T.-M., Wu, C.-E., & Lee, F. Y. (2006). Neighborhood influence on the formation of national identity in Taiwan: Spatial regression with disjoint neighborhoods. *Political Research Quarterly*, *59*, 35—46.

Lipset, S. M. (1959). Some social requisites of democracy. *American Political Science Review*, *53*, 69—105.

Lofdahl, C. (2002). *Environmental impacts of globalization and trade: A systems study*. Cambridge: MIT Press.

Malloy, T. E., & Kenny, D. A. (1986). The social relations model: An integrative method for personality research. *Journal of Personality*, 54, 199—225.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*, 1246—1266.

Moran, P. A. P. (1950a). Notes on continuous stochastic phenomena. *Biometrika*, *37*, 17—23.

Moran, P. A. P. (1950b). A test for serial independence of residuals. *Biometrika*, *37*, 178—181.

Morrow, J. D., Siverson, R. M., & Tabares, T. E. (1998). The political determinants of international trade: The major powers, 1907—90. *American Political Science Review*, *92*, 649—661.

Murdoch, J. C., Sandler, T., & Sargent, K. (1997). A tale of two collectives: Sulfur versus nitrogen oxides emission reduction in Europe. *Economics*, *64*, 281—301.

Ord, J. K. (1975). Estimation methods for models of spatial interactions. *Journal of the American Statistical Association*, *70*, 120—126.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, *27*, 286—306.

Pollins, B. M. (1989a). Conflict, cooperation, and commerce: The effect of international political interactions on bilateral trade flows. *American Journal of Political Science*, *33*, 737—761.

Pollins, B. M. (1989b). Does trade still follow the flag? A model of international diplomacy and commerce. *American Political Science Review*, *83*, 465—480.

R Development Core Team. (2004). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (ISBN 3-900051-00-3; http://www.R-project.org).

Ripley, B. D. (1981). *Spatial statistics*. New York: Wiley.

Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge, UK: Cambridge University Press.

Rose, A. K. (2004). Does the WTO really increase trade? *American Economic Review*, *94*, 98—114.

Rozanski, J., & Yeats, A. (1994). On the (in)accuracy of economic observations: An assessment of trends in the reliability of international trade statistics. *Journal of Development Economics*, *44*, 103—130.

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. London: Chapman & Hall.

Schabenberger, O., & Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Boca Raton, FL: Chapman & Hall.

Shin, M. E. (2001). The politicization of place in Italy. *Political Geography*, *20*, 331—352.

Shin, M. E., & Agnew, J. (2002). The geography of party replacement in Italy, 1987—1996. *Political Geography*, *21*, 221—242.

Shin, M. E., & Agnew, J. (2007a). *Berlusconi's Italy: Where it started, where it ended*. Philadelphia: Temple University Press.

Shin, M. E., & Agnew, J. (2007b). The geographical dynamics of Italian electoral change, 1987—2001. *Electoral Studies*, *26*, 287—302.

Signorino, C., & Ritter, J. (1999). Tau-b or not tau-b. *International Studies Quarterly*, *43*, 115—144.

Tiefelsdorf, M. (1972). The saddlepoint approximation of Moran's I and local Moran's I: Reference distributions and their numerical evaluation. *Geographical Analysis*, *34*, 187—206.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1992). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Varian, H. R. (1972). Benford's law. *American Statistician*, *26*, 65.

Wainer, H. (2004). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, *121*, 311—324.

Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, *92*, 607—617.

Ward, M. D., & Gleditsch, K. S. (2002). Location, location, location: An MCMC approach to modeling the spatial context of war and peace. *Political Analysis*, *10*, 244—260.

Ward, M. D., & Hoff, P. D. (2007). Persistent patterns of international commerce. *Journal of Peace Research*, *44*, 157—175.

Ward, M. D., Siverson, R. M., & Cao, X. (2007). Disputes, democracies, and dependencies: A reexamination of the Kantian peace. *American Journal of Political Science*, *51*, 583—601.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. New York: W. W. Norton.

West, W. J. (2005). Regional cleavages in Turkish politics: An electoral geography of the 1999 and 2000 national elections. *Political Geography*, *24*, 499—523.

You, J.-S., & Khagram, S. (2005). A comparative study of inequality and corruption. *American Sociological Review*, *70*, 136—157.

译名对照表

ad hoc	事先设定的
apparent population	明显的总体
areal data	地区数据
autologistic	自相关逻辑斯蒂
Benford's Law	班佛定律
bias	偏误
Collegi	选区
Conditionally autoregressive model，CAR	条件自回归模型
cross-validation	交叉验证
distribution of first digits	首位数字分布
equilibrium impact	均衡效应
fixed effects	固定效应
Geary's C	吉尔里 C 统计量
Gibbs Sampling	吉布斯抽样
global correlation	全局相关
Imputation	插补方法
inconsistency	不一致
Intercept	截距
inverse	逆矩阵
Jacobian of the transformation	雅克比行列式转换
kernel density	核密度分布
lattice data	晶格数据
Local Indicator of Spatial Association，LISA	局部空间相关指标
map mashups	地图混搭程序
Markov chain Monte Carlo	马尔可夫链蒙特卡罗方法
Maximum likelihood estimator，MLE	最大似然估计量
Mean Squared Errors	均方误
mean-normalized cross-product	均值正态化后得到的内积
metric	度量学
Monte Carlo simulation	蒙特卡洛模拟
Moran I	莫兰 I 统计量

Ord Approach	奥德方法
out-of-sample prediction test	样本外的预测检验
Parsimony	简约性
point data	点状数据
point processes	点过程
pooled OLS	合并数据的 OLS
rug plot	毯图
Shin Plot	锡恩图
Shin spatial scatterplot	锡恩空间散点图
simultaneity	共时性
simultaneous model	联立模型
sparse list	稀疏形式
sparse matrix	稀疏矩阵
spatial autoregressive	空间自回归
Spatial Econometrics	空间计量经济学
Spatial Error	空间性误差
Spatial lag	空间滞后
Spatial Statistics	空间统计学
Spatially Lagged Dependent Variable	空间滞后因变量
super-population	超总体
Time Series Cross Section，TSCS	时间序列的横截面



格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic回归入门
39. 解释概率模型：Logit、Probit以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic回归中的交互效应
46. 应用回归导论
47. 档案数据处理：研究"人生"
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践

上架建议：社会研究方法



微信



微博

ISBN 978-7-5432-2615-9



定价：25.00元

易文网：www.ewen.co

格致网：www.hibooks.cn

[General Information]
书名=空间回归模型
作者=（美）迈克尔?D.沃德（MICHAEL
页数=136
SS号=13976937
DX号=
出版日期=2016.04
出版社=格致出版社；上海人民出版社